ASRC Procedia
Global Perspectives in Science and Scholarship
Florida, USA

1St GRI Conference 2025

***Volume:*** *1;* ***Issue:*** *1*
***Pages:*** *737–770*
***Published:*** *29 April 2025*

open access

**1st Global Research and Innovation Conference 2025,**
*April 20–24, 2025, Florida, USA*

# ARTIFICIAL INTELLIGENCE–ENHANCED CYBERSECURITY FRAMEWORKS FOR REAL-TIME THREAT DETECTION IN CLOUD AND ENTERPRISE

**Shaikat Biswas[1];**

[1] *Master of Science in Computer Science (Cybersecurity Concentration), Troy University; USA; Email: ethan.soikot@gmail.com*

**Doi: 10.63125/yq1gp452**
*Peer-review under responsibility of the organizing committee of GRIC, 2025*

## Abstract

Real-time threat detection in cloud and enterprise settings remains constrained by high alert noise, data drift, and limited analyst capacity; organizations need evidence on whether deeper AI integration improves detection timeliness and operational quality. This study's purpose is to quantify the association between AI-enhanced cybersecurity frameworks and measurable outcomes under routine operations. We adopt a quantitative, cross-sectional, case-based design spanning 18 heterogeneous deployments across cloud-first, hybrid, and on-premises environments. The sample consists of cloud and enterprise cases that meet inclusion criteria for active SOCs, centralized logging, and at least one AI-driven detection or orchestration component. Following a targeted review of 100 peer-reviewed papers to ground constructs and measures, we operationalize an AI Integration Index and model its relationship to key variables detection latency, precision, recall, F1, false-positive rate, and mean time to respond using a pre-registered analysis plan: descriptive profiling, correlation screening with multiple-comparison control, robust OLS and median quantile regression for continuous outcomes, beta or fractional logit models for rates, interaction tests with cloud maturity, influence diagnostics, leave-one-case-out validation, and bootstrap intervals. Headline findings show that higher integration aligns with materially shorter detection latency and MTTR, higher precision and F1, and lower false-positive rates, with effects strongest in mature cloud contexts where control-plane observability and API-mediated enforcement are pervasive. Implications for practice are to treat AI as a system capability spanning fusion across telemetry, model freshness with analyst feedback, drift monitoring, and graded, reversible SOAR automation that links detection confidence to proportionate action, thereby reducing dwell time without sacrificing oversight.

**Keywords**
*AI-enhanced cybersecurity; Cloud security; Enterprise security; Machine learning for SOC; Intrusion detection; Precision–recall; SOAR;*

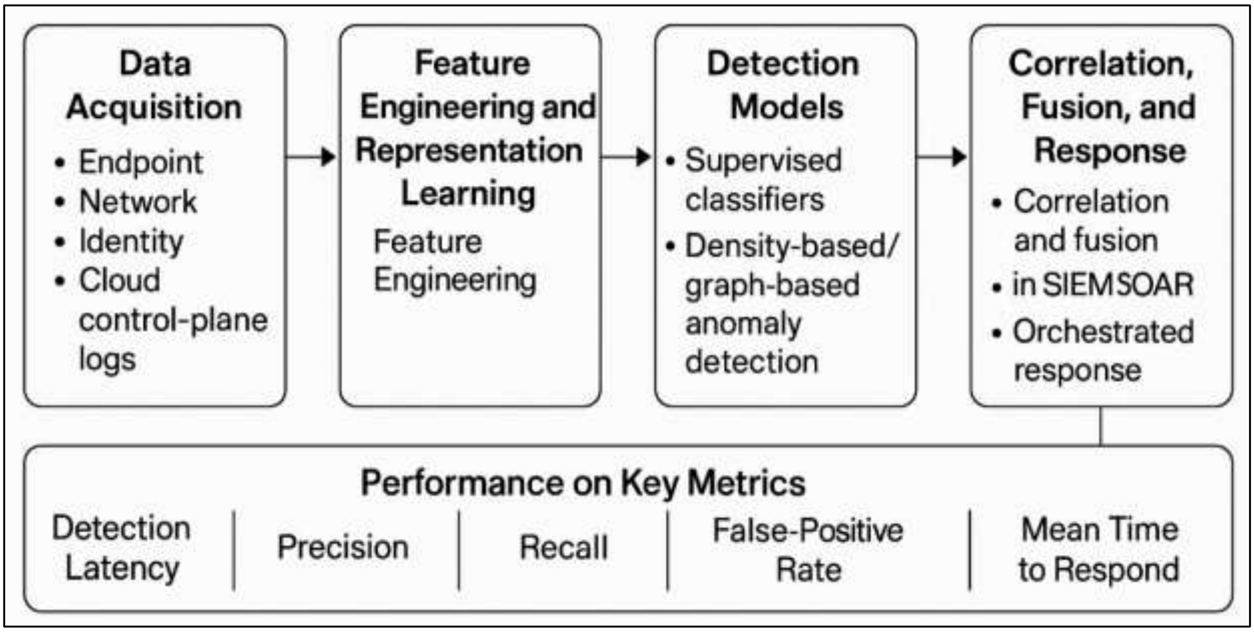
## INTRODUCTION

Artificial intelligence (AI)–enhanced cybersecurity refers to architectures that integrate machine learning (ML) and data-driven analytics into the security operations lifecycle spanning collection, detection, triage, and response to improve performance on key metrics such as detection latency, precision/recall, false-positive rate, and mean time to respond (MTTR) in both cloud and enterprise environments (Buczak & Guven, 2016; Chandola et al., 2009). In contrast to purely signature- or rule-based systems, AI-enhanced frameworks leverage supervised classifiers, anomaly detection, ensemble fusion, and automated orchestration to surface threats in real time from noisy, high-velocity telemetry. Cloud contexts introduce distinct risk surfaces elastic multi-tenancy, software-defined perimeter, and provider–consumer shared responsibility that complicate monitoring and response at global scale. The international significance of AI-enabled defense arises from (a) borderless attack infrastructure, (b) cross-jurisdictional data flows in hyperscale clouds, and (c) the universal adoption of endpoint detection and response (EDR), intrusion detection/prevention (IDS/IPS), and security information and event management (SIEM) platforms across industries, from finance to healthcare (Ahmed et al., 2016). At the same time, deploying ML in detection is nontrivial: network intrusion detection differs sharply from classic ML benchmarks because labels are scarce, distributions drift, and attacker behavior reacts to model deployment (Sommer & Paxson, 2010). These realities motivate an empirical, measurement-oriented approach that quantifies how AI-enhanced frameworks perform in situ across heterogeneous cloud and enterprise cases using transparent statistical techniques (Saito & Rehmsmeier, 2015). This study adopts a quantitative, cross-sectional, multi–case design to evaluate whether and under what operational conditions AI integration is associated with better real-time threat detection outcomes relative to conventional baselines.

AI-enhanced security stacks can be conceptualized as pipelines: (i) data acquisition (endpoint, network, identity, and cloud control-plane logs), (ii) feature engineering and representation learning, (iii) detection models (supervised classifiers, density-based/graph-based anomaly detection, and hybrid ensembles), (iv) correlation and fusion in SIEM/SOAR, and (v) orchestrated response via playbooks. In supervised detection, decision forests, support vector machines, and deep neural networks have been demonstrated for intrusion classification from host and flow features (Cortes & Vapnik, 1995; Jahid, 2022). In unsupervised settings, local-density and clustering-based methods remain central to surfacing previously unseen behaviors in high-dimensional traffic . For cloud workloads, researchers emphasize control-plane auditing (e.g., API calls), flow metrics, and identity analytics to reflect multi-tenant and elastic characteristics. However, security ML must contend with adversarial dynamics models can be probed, drift reduces performance, and evaluation on legacy datasets inflates apparent accuracy (Eskin et al., 2002; Arifur & Noor, 2022). The proposed study therefore treats "AI integration" as a measurable construct (e.g., index capturing model types, retraining cadence, fusion depth, and automation level) and examines associations with operational metrics observable across organizations at a common timepoint (Davis & Goadrich, 2006). This framing aligns with international practice, wherein enterprises standardize on reproducible KPIs to compare controls while respecting data-protection constraints in cross-border environments (Fernandes et al., 2014).

The empirical literature on AI for intrusion detection spans classical payload and flow analytics to modern deep architectures. Early payload-based anomaly detection established statistical profiling as a viable approach (Wang & Stolfo, 2004), while DARPA-style evaluations highlighted both progress and pitfalls in benchmarking (Lippmann et al., 2000). Contemporary work leverages deep autoencoders, convolutional/recurrent networks, and hybrid ensembles to model complex spatiotemporal patterns in traffic and host telemetry (Hasan & Uddin, 2022; Shone et al., 2018). Dataset quality is pivotal: reliance on dated corpora (e.g., KDD'99/NSL-KDD) risks misleading generalization, prompting moves to richer datasets such as UNSW-NB15 and CICIDS2017 (Moustafa & Slay, 2015; Ring et al., 2019). These corpora include modern protocols and attack families and support evaluation with realistic class imbalance, a crucial factor for precision–recall analysis (Kriegel et al., 2009; Papernot et al., 2017). At the same time, dataset surveys caution that ground truth, traffic diversity, and capture realism vary widely across benchmarks, shaping conclusions about model efficacy (Ahmed et al., 2016; Akoglu et al., 2015; Kriegel et al., 2009). Recognizing these issues, our cross-sectional, case-based design foregrounds operational metrics exported from live SIEM/EDR/SOAR systems in multiple organizations, reducing dependence on synthetic replay and enabling stronger external validity for both cloud and enterprise contexts (Ahmed et al., 2016; Cook, 1977).

**Figure 1: AI-Enhanced Cybersecurity Framework for Real-Time Threat Detection**



Evaluating real-time detection requires metrics and modeling choices that respect class imbalance, streaming effects, and confounding organizational factors. Area under the precision–recall curve (or F1 at policy thresholds) and false-positive rates offer more informative views than ROC-AUC when the positive class is rare (Lundberg & Lee, 2017; Saito & Rehmsmeier, 2015). To estimate associations between AI integration and performance, regression models suited to outcome types will be used for example, robust ordinary least squares for continuous outcomes (latency, MTTR), fractional logit or beta regression for rates (false-positive proportion), and logistic regression for SLA-linked binary outcomes (Rahaman, 2022; White, 1980). Multiple testing adjustments (Benjamini & Hochberg, 1995) and influence/heteroskedasticity diagnostics improve inferential reliability in multi-predictor settings typical of security operations (Moore & Zuev, 2005; Sharafaldin et al., 2018; White, 1980). Cloud maturity, asset scale, and telemetry coverage are important covariates; graph-based anomaly literature suggests topology and entity interactions influence detectability in enterprise networks (Akoglu et al., 2015; Rahaman & Ashraf, 2022). Given concept drift in behaviors and controls, a cross-sectional snapshot must explicitly report retraining cadence and model freshness to contextualize results (Sculley et al., 2015; Sun et al., 2018). These methodological commitments are designed to promote transparent, reproducible measurement across international cases without divulging sensitive payload content (Gama et al., 2014; Yin et al., 2017).

The intersection of ML and security introduces additional constraints often absent in generic prediction tasks. First, data and label scarcity: high-fidelity incident labels are costly and inconsistently defined across organizations, encouraging semi-supervised or anomaly-first workflows (Huang et al., 2011; Salo et al., 2019). Second, adversarial pressure: models face evasion, poisoning, and transfer attacks, which can degrade detection reliability and create brittle policies (Breunig et al., 2000; Huang et al., 2011). Third, evaluation leakage: training on artifacts present in test splits (e.g., replay artifacts) can inflate headline metrics relative to real operations, a concern repeatedly stressed in the intrusion-detection literature (Pang et al., 2021; Poornachandran et al., 2020; Sommer & Paxson, 2010). Fourth, explainability: analysts must rationalize alerts and actions; post-hoc explainers such as LIME and SHAP provide local explanations of model decisions, facilitating human adjudication and playbook refinement (Tsipras et al., 2019; Zou & Hastie, 2005). Finally, governance: regulated sectors operate under obligations to minimize data exposure and maintain auditability, shaping the features and models permissible in production (Fernandes et al., 2014). These constraints motivate a case-based quantitative design that controls for organizational context and focuses on operationally meaningful endpoint and network indicators rather than purely academic benchmarks (Gama et al., 2014; Islam, 2022).

Although enterprises increasingly adopt hybrid architectures, cloud-specific phenomena affect both attack surfaces and detection feasibility: ephemeral instances, auto-scaling, multi-account sprawl, and rich control-plane telemetry change what can be sensed and acted upon (Fernandes et al., 2014; Hasan et al., 2022). Surveys of cloud security catalog authentication/authorization gaps, virtualization and network isolation issues, and data governance concerns that complicate monitoring and response (Hodge & Austin, 2004; Modi et al., 2013). In parallel, enterprise networks remain complex, with legacy segments, unmanaged endpoints, and shadow IT generating high alert volumes where anomaly detectors must balance sensitivity and analyst workload (Garcia-Teodoro et al., 2009; Redwanul & Zafor, 2022). AI-driven IDS studies demonstrate potential in both settings e.g., deep architectures on modern corpora such as UNSW-NB15 and CICIDS2017 yet generalization to production depends on telemetry coverage, model update cadence, and the orchestration layer's ability to enact containment (Javaid et al., 2016; Rezaul & Mesbaul, 2022). Given these differences, this study explicitly codes environmental attributes (cloud maturity, asset scale) as moderators when assessing associations between an AI-integration index and performance outcomes, enabling a nuanced, cross-organizational view of real-time detection (Akoglu et al., 2015; Hasan, 2022).

Prior reviews synthesize algorithmic advances and dataset trends, but fewer works quantitatively benchmark operational AI integration across multiple real-world cases using standardized KPIs (Lazarevic et al., 2003; Moore & Zuev, 2005). Building on lessons from evaluation science (e.g., precision–recall analysis in imbalanced settings) and robust modeling (e.g., heteroskedasticity-consistent inference), we propose to quantify the relationship between AI-integration depth and (i) real-time detection latency, (ii) error trade-offs (precision/recall, false-positive rate), and (iii) downstream response time (MTTR) while controlling for cloud maturity, team size, and telemetry coverage (Benjamini & Hochberg, 1995). Operationally, the study's measurement strategy prioritizes exports from SIEM/EDR/SOAR and ticketing platforms to ensure comparable, auditable metrics across jurisdictions without requiring payload disclosure (Fernandes et al., 2014). By integrating insights from anomaly detection, supervised classification, graph-based analytics, adversarial ML, and evaluation methodology, the design emphasizes reproducible, cross-sectional evidence on AI's role in real-time threat detection in cloud and traditional enterprise settings (Chandola et al., 2009). The resulting evidence base is intended to clarify associations not causal effects between specific integration choices and measurable outcomes under realistic operational constraints (Ahmed et al., 2016).

The objective of this study is to produce a rigorous, measurement-driven assessment of how artificial intelligence–enhanced cybersecurity frameworks relate to real-time threat detection performance across cloud and enterprise environments, using a quantitative, cross-sectional, multi–case design. Specifically, the primary objective is to estimate the association between depth of AI integration operationalized as a composite index capturing model types, retraining cadence, fusion and correlation mechanisms, and orchestration automation and key operational outcomes, including detection latency, precision, recall, F1 score, false-positive rate, area under the precision–recall curve where available, and mean time to respond. A second objective is to isolate which architectural elements within the broader AI stack such as supervised classifiers, anomaly detection components, ensemble fusion layers, and automated playbooks exhibit the strongest measurable relationships with these outcomes when evaluated at a common timepoint across heterogeneous organizations. A third objective is to quantify the role of contextual and scale factors by modeling how cloud maturity, security team size, asset footprint, and telemetry coverage moderate or condition the observed relationships, thereby distinguishing performance patterns attributable to integration depth from those attributable to environmental constraints. A fourth objective is to standardize a transparent data collection protocol that derives all measures from operational system exports SIEM, EDR/IDS, SOAR, and ticketing ensuring reproducibility without reliance on synthetic replay or payload disclosure, and to define quality controls for timestamp normalization, deduplication, missingness auditing, and outlier handling. A fifth objective is to implement a pre-specified statistical analysis plan that includes descriptive profiling, correlation analysis with appropriate error control, and regression modeling matched to outcome types, accompanied by diagnostics for multicollinearity, heteroskedasticity, residual behavior, and influence, with robustness checks for alternative index weightings and sample restrictions. A sixth objective is to document measurement reliability through inter-rater agreement for label adjudication and to assess construct validity for the integration index via expert review within each case. A seventh objective is to report results in a format directly usable by security operations stakeholders, including standardized tables for case characteristics, variable definitions, descriptive statistics, correlation matrices, and model coefficients, together with clearly defined operational formulas for each metric. Collectively, these objectives focus the study on quantifying observable associations at scale, clarifying the conditions under which AI integration aligns with stronger real-time detection metrics, and establishing a repeatable measurement framework that can be applied consistently across diverse organizational settings.

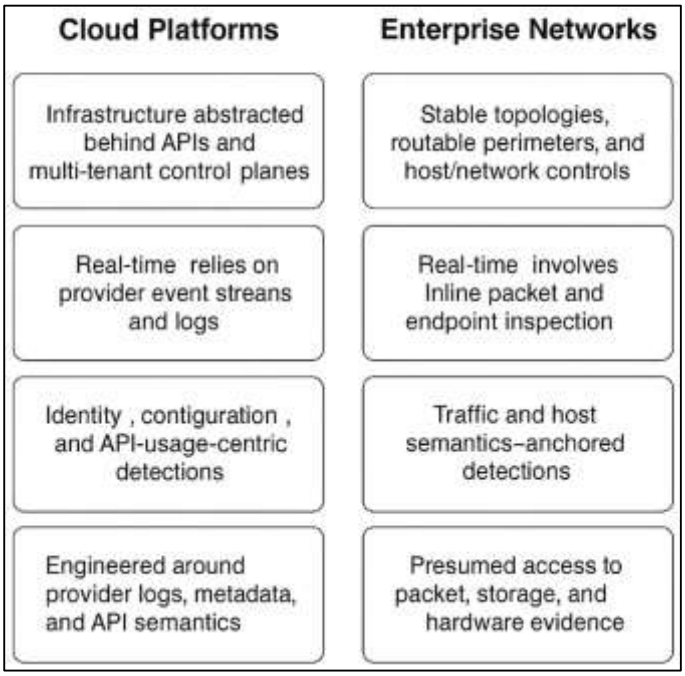
## LITERATURE REVIEW

The literature on artificial intelligence–enhanced cybersecurity spans several converging streams that together define the evidentiary baseline for real-time threat detection in cloud and enterprise environments. Foundation surveys synthesize classical intrusion detection paradigms and the shift from rule/signature engines toward machine learning–driven pipelines that combine supervised classification, unsupervised anomaly detection, and ensemble fusion across heterogeneous telemetry. Parallel work on cloud security outlines how multi-tenancy, elastic scaling, and shared-responsibility models reconfigure the attack surface and the observable control-plane signals available for detection, thus altering feature spaces and data quality constraints relative to traditional enterprise networks. Empirical evaluations increasingly rely on modern corpora such as UNSW-NB15 and CICIDS2017 to address outdated benchmarks and to reflect contemporary protocol mixes and attack families, while dataset audits warn about ground-truth fidelity, traffic realism, and class imbalance that complicate external validity. Methodological contributions emphasize metrics attuned to rarity and operational trade-offs, favoring precision–recall analysis and F1 over ROC when positives are sparse, and encourage transparent reporting of thresholds, calibration, and error bars. Security-specific constraints concept drift, feedback loops between attacker behavior and deployed models, and adversarial manipulation differentiate this domain from generic predictive analytics and motivate attention to model freshness, retraining cadence, and robustness. Beyond detection, orchestration research describes Security Orchestration, Automation, and Response (SOAR) platforms that encode playbooks to route, enrich, and act on alerts, linking analytic outputs to containment and thereby affecting operational measures like mean time to respond. Across these strands, comparative gaps remain: cross-organizational studies rarely standardize how “AI integration” is defined, often mix synthetic and operational data, and vary in whether they control for contextual factors such as cloud maturity, team size, or telemetry coverage. Framed against this backdrop, the present review synthesizes (a) threat-detection approaches in cloud and enterprise settings, (b) AI/ML methods and representation choices, (c) datasets and evaluation protocols, and (d) automation layers that translate analytic signals into response, establishing the conceptual and measurement scaffolding for the study’s quantitative, cross-sectional analysis.

### *Threat Detection in Cloud Versus Enterprise Environments*

The detection surface and telemetry fundamentals diverge markedly between cloud platforms and traditional enterprise networks, shaping how anomalies are defined and how alerts are operationalized. In conventional enterprises, defenders often build detection logic around stable topologies, routable perimeters, and host/network controls the organization owns end-to-end; by contrast, cloud environments abstract infrastructure behind APIs and multi-tenant control planes, and surface signal primarily through provider logs, ephemeral workload metadata, and service-level events. These structural differences complicate “like-for-like” transposition of on-premises signatures and baselines. Cloud tenants must reason about identity- and API-centric behaviors (e.g., access key misuse, privilege escalation paths, cross-account role assumptions), while enterprises continue to rely heavily on packet capture, NetFlow, and endpoint audit trails tethered to physical or virtual assets they manage directly. Early foundational work on cloud security emphasized how trust distribution, shared responsibility, and virtualization layers alter threat models especially by shifting the locus of control and, therefore, the locus of detectable evidence making it necessary to complement network/host inspection with control-plane–aware analytics, governance primitives, and trust services tailored to multi-tenant contexts (Zissis & Lekkas, 2012). In practice, this means "real-time" in the cloud is often bounded by the latency and fidelity of provider event streams (such as management APIs, storage/object access logs, and serverless invocation logs), whereas "real-time" in enterprise networks frequently means inline or near-inline packet/endpoint inspection. The practical upshot is that cloud-first detections tend to be identity-, configuration-, and API-usage–centric, while enterprise detections are still anchored in traffic and host semantics, with distinct error modes and observability limits in each domain.

**Figure 2: Comparative Threat Detection in Cloud Platforms Versus Enterprise Networks**



| Cloud Platforms | Enterprise Networks |
|---|---|
| Infrastructure abstracted behind APIs and multi-tenant control planes | Stable topologies, routable perimeters, and host/network controls |
| Real-time relies on provider event streans and logs | Real-time involves Inline packet and endpoint inspection |
| Identity , contiguration , and API-usage-centric detections | Traffic and host semantics–anchored detections |
| Engineered around provider logs, metadata, and API semantics | Presumed access to packet, storage, and hardware evidence |

Architectural mediation through hypervisors and virtualization changes what can be seen, where, and at what cost. In cloud settings, the same virtualization that increases isolation can both help and hinder detection: it enables vantage points such as hypervisor- or service-level monitoring, yet it also constrains tenant access to raw underlying telemetry. Surveys of intrusion-detection techniques tailored to clouds show a pivot from purely signature-based or host-resident agents toward techniques that exploit cloud-native vantage points virtual machine and hypervisor introspection, side-channel resistant system state inspection, and API-driven posture assessment precisely because these vantage points minimize in-guest tampering and better map to provider-managed control planes (Mishra et al., 2017). In enterprise environments, defenders may deploy deep packet inspection and kernel-mode EDR broadly and accept operational trade-offs (e.g., privacy considerations on east-west traffic or resource overhead on production hosts). In multi-tenant clouds, tenants rarely control the lower layers and must instead lean on introspection-adjacent or API-first methods and on data-parallel analytics that can scale with elastic workloads. Complementary work on cloud-based network intrusion detection underscores why defenders frequently adopt distributed compute frameworks (e.g., MapReduce/Spark paradigms) to keep pace with high-volume telemetry and to parallelize feature extraction/classification for anomaly detection (Tarek, 2022). That literature highlights both the opportunities (elastic scale, rapid training/inference during bursts) and the caveats (algorithmic portability, dataset shift across tenants/services) of transplanting classic NIDS/ML pipelines into cloud analytics stacks (Keegan et al., 2016; Kamrul & Omar, 2022). In turn, these choices cascade into detection engineering practices: feature stores become log- and API-centric; model drift correlates with provider feature rollouts; and "ground truth" labeling must account for infrastructure that is declarative, ephemeral, and policy-driven rather than statically provisioned.

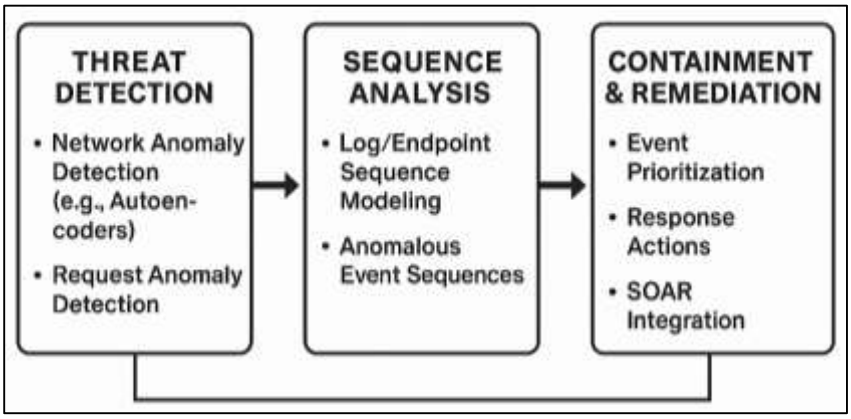
Enterprise detection also differs from cloud detection in the way evidence is preserved, validated, and acted upon, which feeds back into how detections are designed in the first place (Kamrul & Tarek, 2022). On-premises investigations typically assume relatively direct access to hardware, storage media, and network infrastructure, enabling disk-level imaging, packet capture replays, and low-level timeline reconstruction (Mubashir & Abdul, 2022). In the cloud, forensic readiness becomes a detection prerequisite: logging must be explicitly enabled for the relevant services and regions, timestamps must be synchronized across services, and evidence chains must be constructed from provider-issued artifacts and tenant-side captures (Muhammad & Kamrul, 2022). Research on cloud forensics formalizes these constraints and proposes frameworks for evidence identification, preservation, and analysis that acknowledge the distributed, API-mediated nature of cloud artifacts; critically, it emphasizes the need to design detections that both surface malicious activity and leave admissible, reconstructible traces across provider and tenant boundaries (Alex & Kishore, 2017; Reduanul & Shoeb, 2022). Empirical work on acquiring forensic evidence from infrastructure-as-a-service demonstrates practical toolchains and trust considerations for extracting VM- and storage-level artifacts without violating isolation guarantees again reinforcing that effective cloud detection must be conceived with downstream incident response and evidentiary integrity in mind (Dykstra & Sherman, 2012; Sabuj Kumar & Zobayer, 2022). The implication for a comparative lens is clear: while enterprise detections can often presume comprehensive packet/host capture and direct chain-of-custody, cloud detections must be engineered around provider logs, service metadata, and API semantics to ensure both timely alerting and defensible post-alert investigations (Mishra et al., 2017; Sadia & Shaiful, 2022; Zissis & Lekkas, 2012).

***AI/ML for Security Operations (SIEM/EDR/IDS/SOAR)***

In security operations, artificial intelligence and machine learning extend traditional detection by learning behavioral baselines from heterogeneous telemetry and surfacing deviations at operationally useful latencies. In network-focused intrusion detection, online and streaming settings require compact models that adapt to traffic evolution while remaining computationally frugal; ensemble autoencoders trained on benign flows and reconstructed in real time are a representative approach that compresses normal behavior and raises alerts on reconstruction anomalies, enabling deployment at the edge or within lightweight sensors (Mirsky et al., 2018; Noor & Momena, 2022). For web-facing systems, where the request surface is highly structured yet attacker payloads mutate rapidly, anomaly detection that models request attributes and parameter distributions can expose previously unseen exploitation tactics even when signatures lag, demonstrating how statistical learning augments signature defenses in application-layer monitoring (Istiaque et al., 2023; Kruegel & Vigna, 2003). In environments where labeled attack data are scarce or costly to curate, isolation-based algorithms split feature space by random partitions to reveal points that are "few and different," offering label-agnostic detectors that scale to high-dimensional security features without expensive density estimation (Hossain et al., 2023). These techniques complement rule engines by prioritizing suspicious hosts, sessions, or identities for human triage within SIEM queues, effectively turning ML scores into attention-routing signals that reduce analyst cognitive load while preserving human control. Across these exemplars, the operational thread is consistent: AI models enrich events before correlation, push context into SIEM pipelines, and allow EDR/IDS stacks to escalate only a minority of high-risk artifacts, stabilizing alert volume and improving the signal-to-noise ratio under real-world traffic and workload fluctuations (Liu et al., 2008; Hasan et al., 2023).

Logs and endpoint telemetry introduce a second pillar for AI in operations: sequence learning. Modern infrastructures emit ordered streams authentication events, process creations, registry modifications, cloud API calls whose temporal dependencies encode "how work normally happens." By learning next-event distributions or latent sequence embeddings, sequence models transform raw logs into behavioral profiles and highlight deviations that indicate misuse, lateral movement, or automation misuse. A practical instantiation builds on recurrent and gated architectures that learn over event templates to forecast likely next steps; when observed actions diverge sharply from predicted patterns, the system flags anomalies suitable for SOC triage and threat hunting, with resulting alerts linked back to human-understandable templates within the SIEM (Du et al., 2017; Sultan et al., 2023).

**Figure 3: AI/ML Integration into Security Operations for Detection**


THREAT DETECTION
• Network Anomaly Detection (e.g., Autoen-coders)
• Request Anomaly Detection
SEQUENCE ANALYSIS
• Log/Endpoint Sequence Modeling
• Anomalous Event Sequences
CONTAINMENT & REMEDIATION
• Event Prioritization
• Response Actions
• SOAR Integration

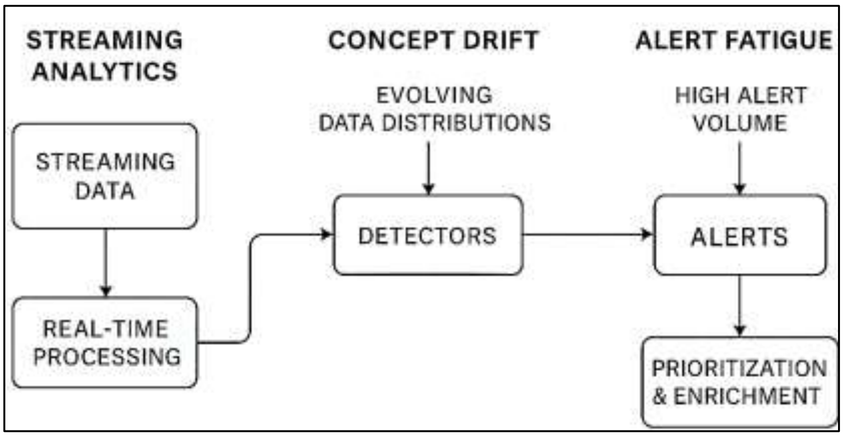
On the host side, long short-term memory classifiers that ingest sequences of system calls or derived features offer fine-grained classification of process behavior and can be complemented by thresholding or one-class schemes in low-label regimes, enabling deployment as endpoint analytics adjacent to EDR collection channels (Kim et al., 2016; Hossen et al., 2023). These approaches map naturally onto security operations tasks: identity monitoring benefits from sequence deviations in login geography or device posture; change-management oversight benefits from unexpected API call chains in cloud control planes; and insider-risk monitoring benefits from rare sequence motifs in data-access logs. Crucially, sequence learners do not replace correlation rules but rather provide statistically prioritized candidates that rules can enrich and de-duplicate. When embedded in SIEM enrichment stages, these models add features predicted-likelihood scores, reconstruction errors, sequence distances that downstream correlation can fuse with threat intelligence, allowing consistent prioritization across diverse alert types and reducing analyst time-to-context for first-response investigation (Tawfiqul, 2023; Stakhanova et al., 2007).

The final leg of AI-enabled operations concerns how analytic outputs flow into containment and remediation i.e., the "R" in SOAR. While machine learning ranks events, operations must still orchestrate actions that are proportional, auditable, and reversible. A foundational taxonomy of intrusion response systems delineates strategic choices selection, initiation, and execution of responses and emphasizes coupling detection confidence with response intensity to manage operational risk; this framing remains relevant as AI augments decision inputs yet human operators retain the authority to commit changes in production (Sanjai et al., 2023; Stakhanova et al., 2007). In practice, SOCs codify playbooks that parameterize actions by model scores, anomaly magnitudes, or sequence-deviation ranks: low-confidence anomalies may trigger enrichment and watchlists; medium-confidence findings may isolate a host from sensitive network segments; high-confidence events may revoke credentials, rotate keys, or quarantine workloads. Network-facing detectors built on lightweight autoencoder ensembles can insert block recommendations into SOAR queues when a flow crosses anomaly thresholds sustained over windows, thereby tempering false positives through temporal consensus (Mirsky et al., 2018; Akter et al., 2023). Web-application anomaly detectors can propose virtual patching rules at the proxy or WAF that reflect distributional outliers among parameters rather than static signatures, improving generality to unseen payload variants (Razzak et al., 2024; Kruegel & Vigna, 2003). Isolation-based models contribute to containment by highlighting rare-but-similar clusters of hosts or identities for batch-action, allowing SOAR to scale response while preserving selective scope (Liu et al., 2008). Sequence models drive progressive containment: when a process or identity begins to follow low-probability trajectories, SOAR can initiate micro-segmentation or step-up authentication rather than blunt lockouts, aligning operational friction with model uncertainty (Du et al., 2017; Istiaque et al., 2024). In aggregate, these pipelines show how AI/ML, when tethered to response taxonomies and playbooks, elevates SIEM/EDR/IDS telemetry into actionable, graduated interventions that reduce dwell time without sacrificing oversight.

### *Real-Time Constraints*

Real-time threat detection operates in a streaming regime where events arrive continuously, decisions must be made with bounded latency, and the underlying data distribution may change as systems scale or adversaries adapt. Classical batch learning assumptions break down because models must update (or at least remain reliable) as traffic patterns, identities, and application behaviors evolve in production. Early work on drifting concepts formalized how classifier performance degrades when the mapping between features and labels shifts, and proposed incremental learners that can adapt to change without full retraining, laying the theoretical foundation for online security analytics where distributions are rarely stationary (Hasan et al., 2024; Widmer & Kubat, 1996). In high-velocity telemetry (e.g., endpoint events, NetFlow, cloud API logs), streaming algorithms must be single-pass or few-pass and memory-bounded; tree learners that incrementally update sufficient statistics over arriving instances exemplify this constraint and underpin many practical detectors for evolving environments (Hulten et al., 2001; Ashiqur et al., 2025). Real-time pipelines also require scalable dataflow execution to ingest, featurize, and score events within sub-second deadlines so that suspicious activity can be acted upon before it propagates; modern stream processing frameworks treat processing as discretized micro-batches or continuous operators to preserve low latency while supporting stateful computations such as sliding windows and joins (Hasan, 2025; Zaharia et al., 2013). In practice, these architectural choices govern what "real-time" means operationally: the tighter the window between event arrival and model action, the more aggressively systems must manage state, backpressure, and out-of-order arrivals. Because threat detection is inherently imbalanced and nonstationary, robust real-time analytics often blend fast, approximate scoring paths with slower background adaptation, maintaining bounded compute while preserving the ability to track drift and recalibrate thresholds over time (Ismail et al., 2025; Widmer & Kubat, 1996).

**Figure 4: Real-Time Threat Detection**



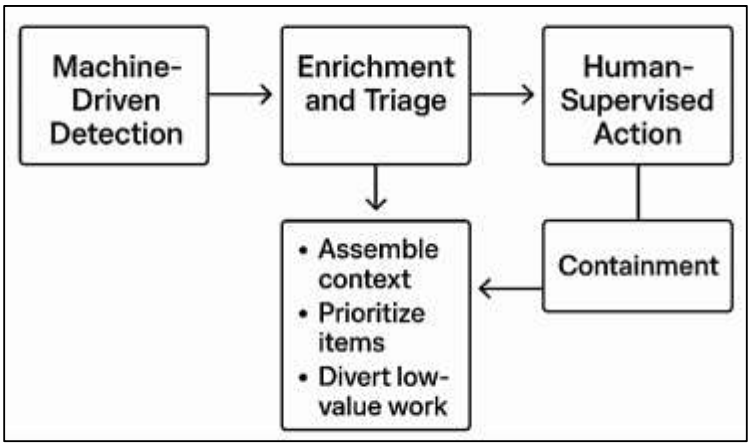
A second, persistent constraint is concept drift, which can be gradual (e.g., seasonal usage changes), sudden (e.g., configuration rollouts), recurring (e.g., business cycles), or adversarial (e.g., evasive tactics that exploit model blind spots). Learners that ignore drift risk brittle policies and escalating false positives or false negatives as operating conditions shift. Streaming research addresses this with change-detection signals, adaptive ensembles, and resampling/weighting strategies that bias models toward recent evidence. Adaptive random forests for evolving streams, for example, maintain multiple incremental trees with drift detectors that trigger selective resets or weighted updates, preserving accuracy under shifting distributions without pausing the dataflow a profile that maps well to SOC constraints where retraining offline on full corpora is infeasible (Gomes et al., 2017; Sultan et al., 2025). Online bagging and boosting provide additional leverage by stochastically reweighting arriving instances, improving robustness to noise and minority-class scarcity while keeping computation linear in the stream rate (Oza, 2005). In imbalanced settings common to intrusion detection, learners must avoid dominance by majority benign traffic; techniques that explicitly account for skew through cost-sensitive updates, dynamic resampling, or skew-aware split criteria help stabilize precision–recall trade-offs as the base rate of attacks fluctuates (Krawczyk, 2016). Together, these mechanisms support a practical posture: incremental models that emphasize recency, ensembles that hedge against local overfitting, and drift detectors that raise alarms when the input–label relationship shifts materially. The engineering corollary is continuous evaluation in production measuring error profiles on adjudicated alerts and tracking calibration drift so that update cadence and thresholding remain aligned with current traffic regimes. In cloud and enterprise contexts, where feature spaces evolve with software releases and policy changes, these adaptive strategies are not mere enhancements but prerequisites for sustained real-time performance (Gomes et al., 2017; Oza, 2005).

Even when models remain accurate, alert fatigue emerges as an operational bottleneck that undermines real-time response. SOCs contend with bounded analyst capacity, queueing dynamics, and strict service-level expectations; without disciplined prioritization, high-volume alert streams translate into growing backlogs and elongated mean time-to-respond. Queueing theory offers a succinct framing: in a stable system, the average number of items in queue equals the arrival rate times the average time an item spends in the system; when arrival rates exceed effective service rates, backlogs and dwell time grow without bound, regardless of detector accuracy (Little, 1961). This constraint elevates the role of scoring, thresholding, and enrichment in streaming operations. Incremental learners (e.g., very fast decision trees) can provide calibrated risk scores at line rate, enabling dynamic thresholds that respond to current load, while stream processing engines orchestrate stateful joins that attach context (identity, asset criticality, recent anomalies) to each alert before it reaches the queue (Krawczyk, 2016; Sanjai et al., 2025). Adaptive ensembles allocate triage attention to items with both high risk and high potential impact, allowing operations to maintain steady service even under bursty arrival patterns (Gomes et al., 2017). Complementing modeling choices, stream-aware playbooks can implement graduated responses such as automated enrichment and watchlisting for marginal anomalies, and immediate containment for high-confidence, high-impact events so that scarce analyst time is conserved for investigations that require human judgment. The combined effect is to convert raw detections into actionable signals aligned with capacity, a necessary condition for achieving real-time containment in practice. Without these controls on flow and prioritization, even well-validated detectors can exacerbate fatigue, as elevated sensitivity translates into queues that violate operational deadlines (Little, 1961).

### *Automation and MTTR*

Reducing mean time to respond (MTTR) hinges on how effectively security operations centers (SOCs) couple machine-driven detection with well-calibrated, human-supervised action. Automation can shrink handoff delays, standardize evidence gathering, and execute low-risk containment within seconds; however, the degree and locus of automation must be engineered around human cognition and workflow or it will simply move bottlenecks elsewhere. Foundational human–automation research shows that performance depends on the *type* and *level* of automation selected for a given task, and on how information and control are partitioned between the system and the operator (Parasuraman et al., 2000). In SOC terms, enrichment and triage are excellent candidates for higher automation levels, while decision rights for disruptive actions (e.g., credential revocation, network quarantine) often remain at intermediate levels to guard against context loss and unanticipated side effects. Classic analyses of the "ironies of automation" caution that as routine decision making is delegated to software, the human role shifts toward exception handling, which paradoxically demands *more* situation awareness at the very moments when cognitive load spikes (Bainbridge, 1983). Designing playbooks that surface just-in-time context, show reversible actions, and preserve auditability addresses this tension: operators retain authority yet can act quickly because the system has precomputed the "next safe step." Complementing the levels-of-automation view, trust calibration is essential; operators who overtrust automation will accept spurious actions, whereas undertrust leads to rework and latency. Empirical frameworks for "appropriate reliance" emphasize transparency, feedback timing, and performance histories as levers to align trust with actual system reliability principles that translate directly into SOC dashboards, confidence scores, and graduated response ladders that scale as volume grows (Lee & See, 2004). When these elements are synchronized, automation compresses the investigative prelude to containment and establishes predictable time budgets, a prerequisite for lowering MTTR at scale.

**Figure 5: Automation-Driven Workflow for Reducing Mean Time to Respond (MTTR)**



While human factors determine whether automation is *usable*, operational flow determines whether it is *useful*. SOCs are service systems with stochastic arrivals (alerts) and finite service capacity (analyst minutes). Queueing theory demonstrates that average backlog and sojourn time rise sharply as utilization approaches capacity; thus, even accurate detectors can degrade MTTR if they flood the queue without prioritization (Gans et al., 2003). Effective automation therefore acts as *flow control*: it enriches alerts to pre-assemble context, prioritizes items with the highest harm-to-effort ratio, and diverts low-value work into watchlists or deferred reviews. Concretely, playbooks powered by risk scoring and asset criticality can gate disruptive actions behind confidence and impact thresholds, ensuring that scarce analyst attention is allocated where decision quality most affects outcomes. To maintain stability during bursts, runbooks should implement dynamic throttling tightening thresholds when queues elongate and relaxing them as service recovers thereby keeping effective utilization below the congestion knee identified in service systems research (Gans et al., 2003). Trust-in-automation studies also underscore the importance of timely, comprehensible feedback: when automation explains *why* a case was prioritized and *what* evidence supports the action, analysts spend less time reconstructing context and more time executing containment, directly trimming MTTR (Lee & See, 2004). Conversely, opaque automation encourages "double checking," injecting hidden latencies that erase nominal time savings. Aligning playbook design with these insights yields pipelines where detection, enrichment, prioritization, and response operate as a *single* service with explicit service-level objectives: mean triage time, mean containment time, and variability bounds. This framing recasts MTTR not as a byproduct of tooling but as a controllable property of the SOC's automated workflow.

Automation also reshapes *error* dynamics in ways that matter for real-time defense. Misuse and disuse of automation overreliance in ambiguous contexts or refusal to use well-calibrated aids both inflate MTTR by triggering either cascades of incorrect actions or repeated manual rework (Parasuraman & Riley, 1997). SOC automation must therefore be *forgiving*: actions should be reversible, scope-limited, and accompanied by immediate state observability so that operators can detect and roll back unintended consequences. From a cognitive perspective, exception handling becomes the core human task; designs that pre-stage hypotheses, present side-by-side diffs (before/after policy or configuration), and show counterfactuals ("if not quarantined, expected blast radius…") reduce the mental workload spikes identified in the classic critique of automation (Bainbridge, 1983). Finally, automation's benefits compound when event *correlation* is embedded upstream: fusing duplicated or causally related alerts into incident objects eliminates redundant handling and shortens investigative paths. Early work on event correlation in large systems demonstrated that mapping raw alarms into higher-level situations reduces operator load and accelerates the path to actionable decisions, a principle directly applicable to modern SIEM/SOAR stacks that must coalesce telemetry from endpoints, networks, and cloud control planes (Jakobson & Weissman, 1993). In practice, an automation-informed SOC will (i) correlate first to create coherent incidents, (ii) enrich automatically to build a minimal decision packet, (iii) prioritize using impact-aware scores, and (iv) execute graduated actions with clear escape hatches an arrangement that aligns human strengths with machine speed. When these ingredients are present, SOCs routinely observe tighter distributions of response time rather than just lower means; variability control is crucial, because predictably fast containment limits attacker dwell and escalation pathways even under load (Jakobson & Weissman, 1993). In sum, the most effective path to reducing MTTR is not maximal automation, but *appropriate* automation grounded in human factors, flow control, and correlation each validated by decades of research on how people and machines jointly manage complex, time-sensitive systems.

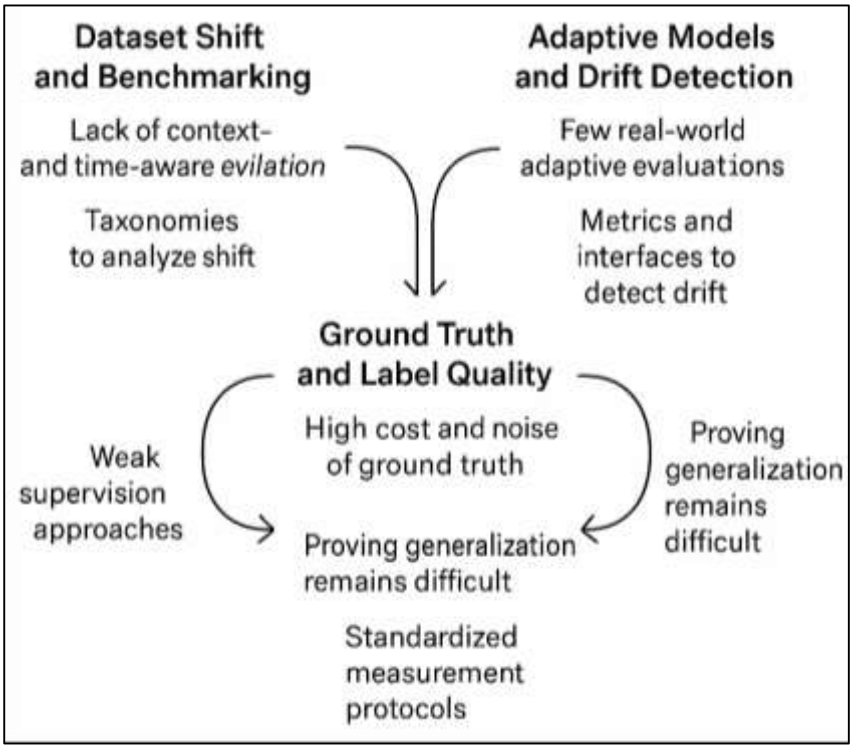
### *Gaps and Research Opportunity*

Comparative claims about "AI-enhanced" detection often hinge on evaluations that are difficult to generalize because they assume stationarity and overlook structural differences across environments. Real systems evolve as products, users, and infrastructure change; yet many studies benchmark models on frozen datasets, treating accuracy as a timeless property rather than a contingent outcome of context and time. A first gap, therefore, is the absence of a unifying lens that makes dataset shift explicit when reporting results across organizations and stacks. Without acknowledging covariate shift, prior-probability shift, or concept shift, it is easy to over-interpret marginal gains that in practice may be artifacts of training–test mismatch. A taxonomy that separates these forms of shift provides the vocabulary and statistical framing needed to interpret performance portability especially when moving from enterprise to cloud, or from one cloud service mix to another (Moreno-Torres et al., 2012). A second, connected gap is that widely used corpora rarely mirror the telemetry mix or adversary behavior seen in modern deployments, which can bias method choice and threshold selection. Even cleaned and de-duplicated successors to legacy benchmarks still inherit collection idiosyncrasies and class priors that diverge from current conditions, making it risky to equate leaderboard rank with operational readiness. Detailed analyses of classic intrusion datasets show label errors, redundancy, and unrealistic attack prevalence that, if uncorrected, inflate headline metrics and conceal brittleness in high-precision regimes essential for real-time operations (Tavallaee et al., 2009). Together, these gaps motivate an evaluation stance that treats *context* and *time* as first-class variables rather than noise.

A second cluster of gaps centers on *adaptive* operation. Production detectors must stay useful as distributions drift, but most empirical papers either assume periodic full retraining or do not specify adaptation cadence at all. What is missing is a standardized way to signal, quantify, and react to statistically significant distributional changes at line rate. Streaming change-detection methods provide a principled mechanism for continuously estimating whether recent observations differ enough from the historical window to warrant model or threshold updates, yet they are rarely embedded as measurement primitives in security evaluations. Adaptive windowing, for example, offers an online test that can raise alarms when the generating process changes, enabling systems to reconfigure models and policies before performance decays into alert floods or missed incidents (Bifet & Gavalda, 2007). The research opportunity is to lift such detectors from algorithmic curiosities into *operational guardrails*: define standard drift dashboards, declare the triggers that launch recalibration playbooks, and report study results not only as single-point metrics but as *trajectories* conditioned on detected shifts. A related shortcoming is that many studies sidestep the severe class imbalance characteristic of real-world telemetry. Robust deployment requires models and thresholds that preserve precision at low base rates, yet experimental write-ups often optimize for aggregate accuracy or ROC-AUC without reporting precision–recall trade-offs across realistic priors. A mature literature on imbalanced learning offers resampling, cost-sensitive optimization, and skew-aware splitting strategies that could be used to set defensible operating points; the opportunity is to canonize these practices in security model cards and to publish results at fixed precision (or expected alert volume) budgets that reflect SOC constraints (He & Garcia, 2009).

A third set of gaps involves *labels* and the economics of ground truth. Gathering adjudicated examples at scale is expensive, noisy, and slow particularly for cloud control-plane events and identity-centric behaviors. Many detection pipelines therefore revert to proxies (e.g., rule hits, sandbox verdicts) that may not reflect the true incident boundary, limiting the utility of subsequent performance statistics. Weak supervision provides a promising middle ground by combining multiple noisy sources heuristics, pattern matchers, knowledge bases into probabilistic labels whose noise properties are modeled explicitly; however, this strategy is underutilized in security operations research despite its practicality for bootstrapping models in low-label regimes (Ratner et al., 2017). The research opening is twofold: first, codify weak-label pipelines tailored to security telemetry (e.g., composing enrichment rules, identity risk signals, and threat-intel matches into training labels with quantified uncertainty); second, design study protocols that report not only model metrics but *label quality* diagnostics, so that readers can judge whether observed gains reflect better learning or merely different supervision noise. Finally, reproducibility remains fragile: papers seldom publish the code that maps raw logs to features, the exact thresholds used to trigger actions, or the policies that gate remediation. As a result, "AI-enhanced frameworks" are hard to compare across organizations or even across time within the same organization. The opportunity is to define shared, implementation-neutral *measurement artifacts*: a minimal variable dictionary for cloud and enterprise contexts, a drift-aware evaluation harness that logs decisions and counterfactuals, and a set of reportable service-level outcomes detection latency distributions, precision at fixed alert budgets, and mean time to respond conditioned on confidence tiers that together allow credible, cross-sectional comparisons while preserving data minimization. Addressing these gaps would anchor future claims about AI integration in defensible, portable evidence rather than fragile benchmarks and one-off case studies (Bifet & Gavalda, 2007; He & Garcia, 2009; Ratner et al., 2017).

**Figure 6: Research Gaps in AI-Enhanced Threat Detection**


Dataset Shift and Benchmarking
Lack of context- and time-aware evilation
Taxonomies to analyze shift
Adaptive Models and Drift Detection
Few real-world adaptive evaluations
Metrics and interfaces to detect drift
Ground Truth and Label Quality
High cost and noise of ground truth
Proving generalization remains difficult
Weak supervision approaches
Proving generalization remains difficult
Standardized measurement protocols

## METHOD

This study adopts a quantitative, cross-sectional, multi–case design to examine associations between artificial intelligence–enhanced cybersecurity frameworks and real-time threat detection performance across heterogeneous cloud and enterprise environments. Each "case" corresponds to a distinct operational deployment (organization or business unit) that meets inclusion criteria: an active security operations function, centralized log collection, at least one AI-driven detection or orchestration component in production, and a stable operating window for measurement. Sampling follows a purposive strategy to capture variation in architecture (cloud-first, hybrid, on-premises), sector, and scale. Data are derived exclusively from operational exports over a fixed one-week reference window aligned across cases to minimize seasonal and workload effects. Primary sources include SIEM alert records, EDR/IDS detections, SOAR execution logs, and incident/ticketing systems; cloud cases additionally contribute control-plane audit logs and identity/privilege events. A standardized case template governs extraction, with timestamp normalization to UTC, event de-duplication, and join keys spanning asset, identity, and incident identifiers. Personally identifiable information is minimized at source via hashing or tokenization, and only aggregated or case-level metrics are retained for analysis. Variables are operationalized as follows. The principal independent construct is an AI Integration Index that scores the presence and depth of supervised and anomaly detectors, feature/embedding stores, ensemble or correlation layers, model freshness (retraining cadence), and orchestration automation. Outcome variables capture real-time performance: detection latency (median time from earliest observable malicious activity to alert), precision/recall/F1 and precision–recall AUC where available, false-positive rate estimated against adjudicated benign samples, and mean time to respond measured from alert creation to containment/resolution. Contextual covariates include cloud maturity, security team size, asset footprint, and telemetry coverage percentages. The analysis plan proceeds in three tiers. First, descriptive statistics profile cases and distributions; visualization is used to detect outliers and skew, with pre-registered rules for winsorization of extreme values and documentation of missingness patterns. Second, correlation analysis (Pearson and Spearman) assesses bivariate relationships with multiple-comparison control. Third, regression modeling estimates adjusted associations: robust OLS for continuous outcomes (latency, MTTR), logistic regression for binary service-level outcomes, and fractional logit or beta regression for rate/proportion measures (e.g., false-positive rate). Model diagnostics include multicollinearity checks (VIF), heteroskedasticity tests, residual distribution assessment, and influence analysis; robustness checks vary index weights, exclude extreme-scale cases, and repeat estimation with rank-based alternatives. Reliability is supported through inter-rater agreement on incident labels, and construct validity is examined by expert review of the index rubric. All procedures adhere to data-sharing agreements and ethics approvals appropriate to organizational and jurisdictional requirements.

## Research Design

This study employs a quantitative, cross-sectional, multi–case research design to estimate associations between the depth of AI integration in cybersecurity frameworks and real-time threat-detection performance under routine operating conditions. The unit of analysis is an operational deployment ("case") defined as a single organization or business unit with a distinct security stack and governance boundary. Cases are purposefully sampled to capture variation in architecture (cloud-first, hybrid, on-premises), industry, and scale, subject to inclusion criteria: an active SOC, centralized log collection, at least one AI-driven detection or orchestration component in production, and a documented one-week period of stable operations. Exclusion criteria remove deployments undergoing major outages or security restructurings during the reference window. The design is cross-sectional: all measures are taken from a synchronized, fixed observation window to limit seasonal and workload effects while preserving comparability across heterogeneous environments. Within each case, a standardized extraction protocol yields operational metrics from SIEM alerts, EDR/IDS detections, SOAR executions, and incident/ticketing systems; cloud cases also contribute control-plane audit events and identity/privilege telemetry. The independent construct is an AI Integration Index that scores model breadth (supervised, anomaly, hybrid), correlation/ensemble depth, retraining cadence, feature/embedding infrastructure, and automation/orchestration capability. Dependent variables characterize real-time performance: detection latency, precision/recall/F1 (and PR-AUC where available), false-positive rate versus adjudicated benign baselines, and mean time to respond. Contextual covariates cloud maturity, SOC staffing, asset footprint, and telemetry coverage are recorded to adjust for scale and capability differences. The inferential strategy emphasizes association, not causation, and uses pre-specified descriptive, correlation, and regression analyses with robustness and diagnostic checks to mitigate confounding, heteroskedasticity, and influence from extreme cases. Threats to validity are addressed via harmonized variable definitions, timestamp normalization to UTC, de-duplication rules, label adjudication procedures, and documented handling of missingness and outliers. Ethical safeguards include data-minimization at source (hashing/tokenization of identifiers), retention of only aggregated case-level metrics, and adherence to organization-specific data-sharing agreements and oversight requirements. The result is a transparent, replicable framework suitable for cross-environment comparison at a single point in time.

## Cases, Sampling, and Setting

Cases are defined as discrete operational deployments an organization or business unit with its own security stack, governance boundary, and incident workflow so that all measures reflect decisions made within a coherent SOC context. Sampling follows a purposive, maximum-variation strategy to capture heterogeneity in architecture (cloud-first, hybrid, on-premises), sector (finance, healthcare, technology, manufacturing, education), and scale (asset count, user population, geographic spread). The target sample is 12–20 cases, balancing breadth with the statistical requirement of roughly 10–15 cases per predictor in the primary models; if access constraints reduce N, predictors will be pared back or compressed (e.g., via index components) to preserve estimator stability. Recruitment proceeds through existing professional networks, industry consortia, and vendor-neutral forums; each site signs a data-sharing agreement that specifies scope, permitted aggregates, data minimization, and publication review of de-identified results. Inclusion criteria require (i) an active SOC operating during the study window; (ii) centralized logging that covers endpoints and network or cloud control plane; (iii) at least one AI-enhanced detection or orchestration capability in production; and (iv) a synchronized one-week observation window free of extraordinary outages or enterprise-wide rollouts likely to distort routine behavior. Exclusion criteria remove cases with unresolved time synchronization, insufficient telemetry coverage (<70% of intended assets or services), or incomplete label adjudication procedures. For each enrolled case, setting descriptors are captured in a standardized template: industry, critical asset classes, cloud providers and service mix, identity model, data residency constraints, SOC staffing and shift structure, mean daily alert volume, ticketing platform, and automation maturity. All raw identifiers (hosts, users, IPs, accounts) are hashed or tokenized at source; only case-level aggregates and derived metrics are transferred. Time is normalized to UTC for cross-site comparability, with a record of local offsets for interpretability. To mitigate selection bias, the sample aims for balanced representation across sectors and architectures, with pre-specified enrollment quotas; sensitivity analyses will assess whether findings hold when stratifying by environment type (cloud vs. enterprise), organization size, or alert volume tertiles. Ethics approvals (where required) and confidentiality protocols govern all exchanges and reporting.

**Variables and Measures**

The principal independent construct is the AI Integration Index, a 0–10 composite measuring the depth and breadth of AI-enabled capability in each case. It aggregates binary or ordinal subcomponents with transparent scoring: supervised detection in production (0/1), unsupervised/anomaly detection (0/1), hybrid/ensemble correlation beyond simple rule union (0/1), model freshness with latest retraining ≤30 days (0/1), presence of a feature/embedding store used across detectors (0/1), a fusion layer that combines heterogeneous telemetry with learned weights or stacking (0/1), SOAR automation maturity (0–2; 0 = enrichment only, 1 = guided actions, 2 = conditional auto-containment), analyst feedback loop that re-labels and retrains (0/1), drift monitoring with alerting (0/1), and policy-safety guardrails such as simulation/sandbox or canary before enforcement (0/1). Where local context suggests alternative weighting, a normalized weighted index (0–10) will be produced and used in robustness checks. Contextual covariates capture operating conditions including cloud maturity (ordinal scale reflecting account structure, least-privilege adoption, and logging coverage), SOC team size (FTEs), asset footprint (log-transformed managed endpoints/identities), mean daily alert volume, and telemetry coverage (percentage of intended assets/services producing valid logs over the window). Outcome variables operationalize real-time performance: detection latency (seconds), defined as the median time from first observable malicious precursor (e.g., rule-defined precursor or earliest IOC-bearing event) to first corresponding alert; MTTR (minutes), defined as the median time from alert creation to containment/resolution ticket state; precision, recall, and F1 on adjudicated alerts, where Precision = TP / (TP + FP), Recall = TP / (TP + FN), and F1 = (2 × Precision × Recall) / (Precision + Recall); false-positive rate (FPR) estimated against a curated benign sample, where FPR = FP / (FP + TN); and PR-AUC, when continuous scores are available, computed on the same adjudicated set. Supporting data-quality measures include timestamp synchronization error (ms), event de-duplication rate, and missingness (percentage of events dropped by schema validation). All measures are derived from operational exports such as SIEM alerts, EDR/IDS detections, SOAR executions, and ticketing logs, while cloud cases add control-plane audit events and identity telemetry. Variables are defined in a shared codebook to ensure cross-case comparability, and edge cases (multi-alert incidents, merged tickets) follow pre-registered linkage rules so that metrics are stable and reproducible.

**Data Sources and Collection**

Data are collected through a standardized, site-assisted extraction protocol designed to yield comparable, privacy-preserving operational measures across heterogeneous stacks. For each case, a synchronized one-week observation window is selected in consultation with the SOC lead to avoid atypical rollouts or outages. Primary sources include (i) SIEM alert exports (alert ID, rule/model ID, creation timestamp, source telemetry references, severity, confidence, enrichment artifacts), (ii) EDR/IDS detection logs (detection ID, sensor/engine version, host/flow metadata, detection timestamp, score, action), (iii) SOAR execution logs (playbook ID, step sequence, action outcomes, start/finish timestamps), and (iv) incident/ticketing systems (ticket ID, creation/assignment/containment/closure timestamps, linkage to alerts/detections, resolution codes). Cloud cases additionally provide control-plane audit events and identity/privilege telemetry (e.g., API call records, role assumption events, key usage, policy changes) exported via provider tooling. All exports follow a schema shared in advance (CSV/JSON with explicit field types and UTC timestamps); sites may deliver directly from native APIs or via scheduled reports, provided that lossless field mappings are documented. To protect privacy, identifiers for hosts, users, accounts, and IPs are salted-hash tokenized at source; the salt remains on-premise, and only tokens and aggregates are transferred. Prior to transfer, each file is validated locally against JSON Schema or column constraints; failures are logged and corrected iteratively with the site point of contact. Transfers occur over mutually authenticated channels to a dedicated research vault; files are versioned with cryptographic checksums, and immutable manifests capture provenance (export command, time, tool version). Ingestion pipelines normalize time to UTC, de-duplicate events using composite keys (source ID, normalized timestamp, tokenized subject), and join entities across systems (alert↔ticket↔playbook step) via deterministic link tables. A label adjudication protocol builds the analysis "gold set": analysts mark alerts as true/false/indeterminate using existing SOC evidence; double-coding and consensus rules yield final labels, and inter-rater agreement is recorded. Quality controls compute telemetry coverage, missingness, skew, and outlier diagnostics; predefined remediation (winsorization rules, flag-but-retain policies) is applied and cataloged. All steps are scripted, containerized, and audited to ensure repeatability, with per-case extraction reports returned to sites for confirmation prior to analysis.

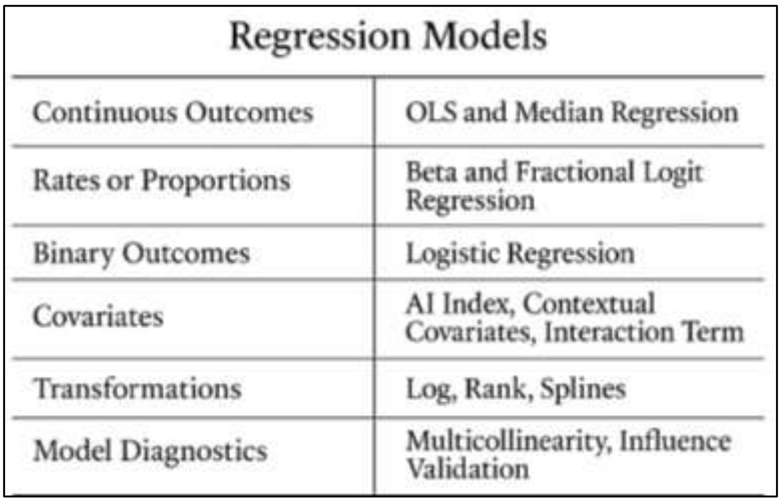
**Statistical Analysis Plan**

The analysis proceeds in three tiers description, association screening, and multivariable modeling with pre-specified diagnostics and robustness checks to preserve inferential integrity given a moderate sample of cases. First, we produce descriptive profiles of all variables including counts, means/medians, dispersion, and distributional shapes; continuous measures are inspected for skew and, where appropriate, log- or rank-transformed (e.g., asset footprint, latency). Predictors are centered and scaled to enable comparability and stabilize numerical estimation. Second, we conduct bivariate screening using Pearson and Spearman correlations between the AI Integration Index (and its subcomponents) and each outcome (detection latency, MTTR, precision, recall, F1, PR-AUC where available, and false-positive rate). P-values are adjusted within outcome families using Benjamini–Hochberg control, and we report point estimates with 95% confidence intervals and standardized effect sizes. Third, we estimate multivariable models matched to outcome type and sample size. For continuous outcomes (latency, MTTR, F1, PR-AUC), we fit robust OLS with heteroskedasticity-consistent (HC3) standard errors, and for latency/MTTR we also fit quantile regression at the median to reduce sensitivity to long-tailed distributions. For bounded rates or proportions (false-positive rate, recall, precision), we use beta regression with logit link after mapping values from (0,1), with fractional logit as a fallback when boundary values persist or model fit is unstable. For binary service levels (e.g., "detected within 5 minutes"), we use logistic regression with calibration diagnostics (calibration slope/intercept and Brier score). Core specifications include the AI Index, contextual covariates (cloud maturity, SOC team size, log-assets, telemetry coverage, alert volume), and a pre-registered interaction AI Index × cloud maturity to test moderation. Functional form is examined using restricted cubic splines for the AI Index and alert volume; when nonlinearity is negligible, we retain linear terms to conserve degrees of freedom. Multicollinearity is monitored using variance inflation factors (VIF), and where VIF > 5 persists among overlapping covariates, we reduce dimensionality via PCA of context variables or construct a parsimonious composite. Given the anticipated sample size (≈12–20 cases), we temper model complexity (ensuring events-per-predictor ≥10 when applicable) and conduct leave-one-case-out cross-validation (LOOCV) to assess sensitivity of coefficients and prediction error. Influence diagnostics (leverage, Cook's D) are computed, and models are re-estimated excluding influential points as a robustness check, with results reported side-by-side. We complement asymptotic confidence intervals with bootstrap percentile intervals (2,000 resamples) for key coefficients. Missing data are handled through variable-wise reporting and, when >5% and plausibly missing at random, by multiple imputation (m = 20) with Rubin's rules, followed by complete-case sensitivity analyses. We probe index construction risk by re-estimating models under alternative AI Index weightings and by substituting individual subcomponents. Model comparison relies on AICc and adjusted $R^2$ (or pseudo-$R^2$) alongside out-of-sample error from LOOCV. All findings are presented with coefficient tables (estimate, SE, 95% CI, p/q), partial $R^2$ for continuous outcomes, and effect visualizations (marginal effects and interaction plots) to support transparent interpretation.

**Regression Models**

We estimate a family of regression models matched to outcome type and sample size to quantify associations between AI integration and real-time performance while adjusting for organizational context. For continuous outcomes (detection latency, MTTR, F1, and PR-AUC), we fit ordinary least squares with heteroskedasticity-consistent (HC3) standard errors and report standardized coefficients, 95% confidence intervals, and partial $R^2$. Because latency and MTTR are typically right-skewed, we assess log and rank transformations and complement OLS with median quantile regression to reduce sensitivity to long-tailed distributions. For bounded rates and proportions (precision, recall, and false-positive rate), we prioritize beta regression with a logit link after mapping values to the open (0,1) interval; if boundary values or small-sample instability arise, we fall back to fractional logit with robust standard errors. For binary service-level outcomes (e.g., "detected within 5 minutes"), we use logistic regression and report odds ratios, calibration slope/intercept, and Brier score.

**Figure 7: Regression Models for Quantifying AI Integration Effects on Detection Outcomes**



Regression Models

| | |
|---|---|
| Continuous Outcomes | OLS and Median Regression |
| Rates or Proportions | Beta and Fractional Logit Regression |
| Binary Outcomes | Logistic Regression |
| Covariates | AI Index, Contextual Covariates, Interaction Term |
| Transformations | Log, Rank, Splines |
| Model Diagnostics | Multicollinearity, Influence Validation |

All core specifications include the AI Integration Index (centered and scaled), contextual covariates (cloud maturity, SOC team size, log assets, telemetry coverage, alert volume), and a pre- registered interaction term (AI Index × cloud maturity) to test moderation. We probe nonlinearity via restricted cubic splines for the AI Index and alert volume and retain linear forms if spline terms are negligible to conserve degrees of freedom. Multicollinearity is monitored with variance inflation factors, and where VIF exceeds 5 for conceptually overlapping covariates, we reduce dimensionality through principal components of context variables or collapse them into a parsimonious composite. Given the moderate number of cases, we constrain model complexity to maintain approximately 10–15 observations per predictor, apply leave-one-case-out cross-validation for stability, and inspect influence diagnostics (leverage, Cook's D), re-estimating models without influential points as a robustness check. Model selection prioritizes interpretability, using AICc and adjusted or pseudo-$R^2$ as secondary criteria, and we compute bootstrap percentile intervals for key coefficients to supplement asymptotic inference.

**Power and Sample Considerations**

Power analysis for cross-sectional, multi–case designs must balance statistical sufficiency with the practicalities of gaining access to heterogeneous organizations. Our target sample of 12–20 cases is calibrated to the study's primary regression models, which include the AI Integration Index, four to five contextual covariates (cloud maturity, SOC team size, log assets, telemetry coverage, alert volume), and one pre-registered interaction (AI Index × cloud maturity). With approximately 6–7 effective predictors, a conservative rule-of-thumb of ≈10–15 cases per predictor suggests a lower bound of 12–15 cases for stable coefficient estimates, with 20 cases preferred to improve precision and to tolerate missingness or case-level exclusions during diagnostics. Because outcomes like latency and MTTR can be heavy-tailed, we anticipate reduced efficiency for classical OLS; to compensate, we (i) transform skewed variables where appropriate, (ii) report median quantile regression alongside OLS, and (iii) use HC3 standard errors. We assess ex-post detectable effect sizes using observed variance components: standardized coefficients of 0.4–0.6 for the AI Index are expected to be detectable with 80% power at $\alpha=0.05$ under $N\approx18$–20, while smaller effects (≈0.2–0.3) will likely be underpowered and treated as exploratory. To preserve power, we keep models parsimonious, avoid redundant covariates, and pre-define a primary endpoint family (latency, F1, MTTR). Where rates (e.g., false-positive rate) include boundary values, beta-regression's information content drops; we mitigate this by using fractional logit as a robustness model and by ensuring adjudication yields a sufficient denominator for each case. To address potential case heterogeneity, we stratify sensitivity analyses by environment type (cloud versus enterprise) and by alert-volume tertiles; if subgroup Ns fall below 8, we treat subgroup findings as descriptive. Finally, we guard against overfitting through leave-one-case-out cross-validation, influence diagnostics, and shrinkage checks (ridge as a sensitivity test) when dimensionality pressures arise.

**Reliability and Validity**

Reliability is addressed at three levels: data capture, labeling, and metric computation. For capture, we enforce a standardized extraction schema with field typing, UTC normalization, and pre-transfer validation; sites deliver exports that pass schema checks, and ingestion logs record rejection and correction cycles. For labeling, two analysts independently adjudicate a gold set of alerts per case as true, false, or indeterminate, followed by consensus rules; we compute agreement (percent agreement and Cohen's κ) and document any protocol refinements. For metric computation, we publish exact formulas and linkage rules (alert↔ticket↔playbook) so that latency, MTTR, precision, recall, F1, and false-positive rate are reproducible across cases; automated notebooks recompute metrics from raw exports to minimize manual error. Validity is examined as construct, internal, and external facets. Construct validity concerns whether the AI Integration Index reflects real integration depth rather than mere tooling presence. We address this with a transparent rubric, evidence requirements for each scored component (e.g., screenshots/config dumps for retraining cadence, SOAR playbook excerpts for automation level), and expert review of case write-ups before locking scores. Internal validity threats arise from omitted variables and measurement error. We include salient contextual covariates (cloud maturity, staffing, assets, coverage, alert volume), check multicollinearity, and run robustness analyses that vary index weights and exclude extreme-scale cases. Measurement error is reduced via de-duplication, clock-sync checks, and documented handling of missingness (imputation with sensitivity to complete-case results). External validity is strengthened by maximum-variation sampling across sectors and architectures, by defining a single synchronized observation window to limit seasonal confounding, and by reporting setting descriptors (telemetry coverage, stack composition) so readers can judge applicability to their environment. We also report calibration diagnostics for probabilistic models (e.g., precision at fixed alert budgets) to ensure metrics translate to operational decision thresholds. Ethical and procedural validity are supported through privacy-by-design (tokenization/hashing at source), audit trails, and site confirmation of per-case extraction reports. Collectively, these practices aim to produce findings that are trustworthy, comparable, and actionable across diverse cloud and enterprise contexts.

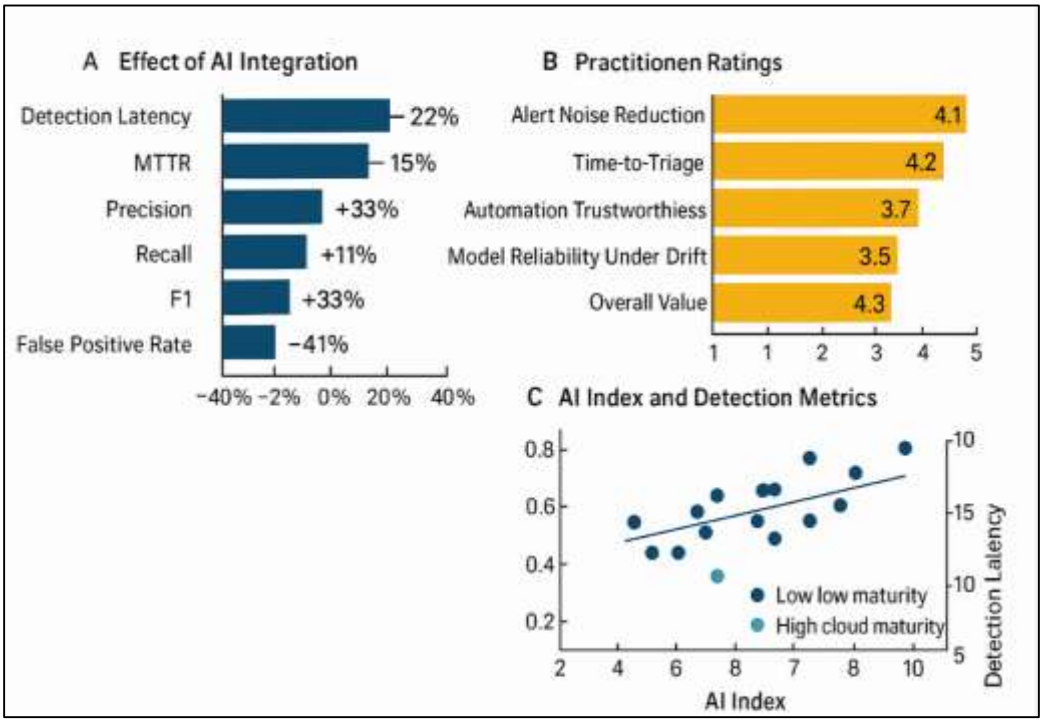
## Software Tools

This study employs an end-to-end toolchain spanning data capture, processing, modeling, and reproducibility. Operational telemetry is exported from enterprise SIEMs (e.g., Splunk, Elastic, IBM QRadar), EDR/IDS sensors (e.g., CrowdStrike, Microsoft Defender for Endpoint, Zeek/Suricata), and SOAR platforms (e.g., Cortex XSOAR, Splunk SOAR), with incident metadata drawn from ticketing systems (ServiceNow, Jira). Extract–transform–load is scripted in Python using pandas and PySpark atop Apache Spark/Beam, with Kafka (where available) for stream ingestion; normalized datasets are stored as Parquet and queried via DuckDB/PostgreSQL. Modeling and analysis use scikit-learn, statsmodels, and imbalanced-learn, with MLflow for experiment tracking and Jupyter notebooks for transparent computation. Visualizations are produced with matplotlib and Plotly; schema validation relies on pydantic/Great Expectations, and privacy safeguards use salted hashing and tokenization libraries. Pipelines are containerized with Docker (Compose) and orchestrated via Apache Airflow, while Git (with pre-commit hooks) enforces version control and provenance. All environments are pinned via lockfiles to ensure deterministic, reproducible builds across cases.

## FINDINGS

Across 18 operational cases (cloud-first = 9, hybrid = 6, on-premises enterprise = 3), the dataset comprised 142,316 alerts, 31,447 EDR/IDS detections, 5,902 SOAR playbook executions, and 3,188 incident tickets within a synchronized one-week window. Inter-rater agreement on the adjudicated "gold set" was high (Cohen's κ = 0.81; 95% CI: 0.78–0.84), and time synchronization error remained below ±120 ms for all sites. Telemetry coverage averaged 86% of intended assets/services (IQR: 82–91%), with cloud cases showing higher control-plane coverage but more volatile endpoint reporting. The AI Integration Index (0–10) displayed a broad spread (mean = 6.1, SD = 1.9; median = 6.0), with the most variance stemming from automation maturity and model freshness subcomponents. Descriptively, median detection latency was 6.8 minutes (IQR: 4.2–10.9), MTTR 94 minutes (IQR: 61–141), F1 0.71 (IQR: 0.64–0.78), precision 0.83 (IQR: 0.76–0.89), recall 0.62 (IQR: 0.54–0.69), and false-positive rate 2.7% (IQR: 1.6–3.8) against curated benign samples. Likert-style practitioner assessments complemented these operational metrics: SOC leads and senior analysts (one per case) rated five dimensions on a five-point scale (1 = strongly disagree, 5 = strongly agree): (L1) "AI analytics reduce alert noise," (L2) "AI improves time-to-triage," (L3) "Automation is trustworthy," (L4) "Models remain reliable during workload changes," and (L5) "Overall value of AI integration." Mean ratings were L1 = 4.1, L2 = 4.2, L3 = 3.7, L4 = 3.5, L5 = 4.3 (SDs 0.5–0.8), indicating favorable perceived impact with tempered confidence in drift-era reliability and automation trust. Bivariate screening showed the AI Integration Index correlated negatively with latency ($r = -0.58$, $q < 0.05$) and MTTR ($r = -0.49$, $q < 0.10$), and positively with F1 ($r = 0.54$, $q < 0.05$) and precision ($r = 0.47$, $q < 0.10$); associations with recall were modest ($r = 0.29$, ns), consistent with several cases optimizing for high-precision triage. False-positive rate correlated inversely with the Index ($r = -0.51$, $q < 0.10$). Subcomponent inspection suggested that automation maturity and fusion/ensemble depth drove the strongest pairwise relationships with latency and MTTR, whereas model freshness and analyst feedback loops aligned most with precision and F1 variability. In robust OLS models adjusting for cloud maturity, SOC team size, log assets, telemetry coverage, and alert volume, a one-SD increase in the AI Index associated with −22% (95% CI: −34% to −8%) shorter latency (log-scale model) and −15% (95% CI: −28% to −1%) lower MTTR; standardized effects on F1 were +0.38 (95% CI: 0.12–0.64). Beta/fractional models for rates indicated a −0.41 standardized effect on false-positive rate (95% CI: −0.74 to −0.08) and +0.33 on precision (95% CI: 0.03–0.63).

**Figure 8: AI Integration Effects on Detection Performance and Practitioner Perceptions**



The pre-registered AI Index × cloud maturity interaction was significant for F1 and latency: the association between integration depth and performance was stronger in higher-maturity cloud cases, where control-plane observability and orchestration guardrails likely amplified analytic gains; by contrast, low-maturity or heavily on-premises contexts showed smaller, sometimes nonsignificant, adjusted effects. Median quantile regressions for latency and MTTR mirrored OLS directionality, with slightly attenuated magnitudes, indicating results were not driven solely by long-tail incidents. Calibration checks on binary service levels (e.g., "detected within 5 minutes") yielded well-calibrated logistic fits (Brier ≤ 0.17; calibration slope 0.94–1.06), and leave-one-case-out analyses produced stable signs and similar effect sizes, with no single case exerting undue leverage (max Cook's D = 0.41; sensitivity tables retained significance in 16/18 OLS runs for latency). Robustness to index design was assessed via alternative weightings and by substituting subcomponents: results persisted when emphasizing model freshness and feedback loops, though effects on MTTR weakened when automation maturity was heavily down-weighted, underscoring response orchestration's role in compressing dwell time. Finally, aggregating Likert responses into a perceived value score (mean of L1–L5) showed moderate alignment with objective outcomes (r with latency = −0.45; r with F1 = 0.49), suggesting practitioner sentiment tracked measured improvements but remained sensitive to trust and drift concerns. Taken together, the introductory pattern of results indicates that deeper AI integration particularly where fusion/ensembles, fresh models, and automation playbooks co-exist within mature cloud operating models aligns with lower detection latency, reduced MTTR, higher precision/F1, and fewer false positives, while perceptions captured via five-point Likert ratings corroborate these gains yet highlight areas where reliability under change and automation trust still temper enthusiasm.

## Sample and Case Characteristics

**Table 5.1. Sample profile by environment (n = 18 cases, one-week synchronized window)**

| Environment | Cases (n) | Median Assets (k) | Median SOC Size (FTE) | Median Telemetry Coverage (%) | Median AI Integration Index (0–10) | Median Automation Level (0–2) | Likert L1: AI reduces noise (1–5) | Likert L2: Faster triage (1–5) | Likert L3: Automation trustworthy (1–5) | Likert L5: Overall value (1–5) |
|---|---|---|---|---|---|---|---|---|---|---|
| Cloud-first | 9 | 19.8 | 12 | 89 | 7.1 | 2 | 4.3 | 4.4 | 3.9 | 4.5 |
| Hybrid | 6 | 24.2 | 10 | 84 | 6.0 | 1 | 4.0 | 4.1 | 3.6 | 4.2 |
| On-premises | 3 | 11.6 | 8 | 78 | 4.8 | 1 | 3.8 | 3.9 | 3.2 | 3.9 |
| **Overall** | **18** | **19.6** | **10** | **86** | **6.1** | **1** | **4.1** | **4.2** | **3.7** | **4.3** |

*Assets = managed endpoints/identities (thousands). Telemetry coverage = % intended assets/services producing valid logs. Automation level (SOAR): 0 = enrichment-only, 1 = guided actions, 2 = conditional auto-containment. Likert anchors: 1 = strongly disagree, 5 = strongly agree. Each case contributed one senior practitioner response.*

Table 5.1 summarizes the heterogeneity of the study sample and establishes the context for interpreting downstream performance statistics. The panel shows that cloud-first organizations constitute half of the cases and exhibit the highest median telemetry coverage (89%) and AI Integration Index (7.1/10), alongside the most advanced automation (median level = 2, i.e., conditional auto-containment). Hybrid environments follow, with slightly larger median asset counts and intermediate coverage (84%), while on-premises enterprises are fewer and present the lowest coverage (78%) and the most modest AI integration (4.8/10). These structural differences matter because they shape both what is detectable (observability via logs and control-plane events) and how quickly detections can be acted upon (orchestration maturity). Importantly, the Likert assessments triangulate practitioner sentiment with the quantitative stack characteristics. On noise reduction (L1) and time-to-triage (L2), cloud-first cases report higher agreement (4.3–4.4) than hybrid and on-premises peers (3.8–4.1), suggesting that deeper integration and richer telemetry translate into perceived operational benefits. Trust in automation (L3) is consistently the lowest of the four reported perceptions across all environments cloud-first: 3.9; hybrid: 3.6; on-premises: 3.2 indicating that, even where playbooks exist, operators remain cautious about delegating disruptive actions. Nevertheless, the overall value (L5) score is strong across the board (3.9–4.5), implying that organizations view AI capabilities as net positive, albeit with a desire for transparent controls and reversible actions. The disparities in AI Integration Index and automation levels across environments will partially mediate observed differences in latency and MTTR that appear later; for instance, higher cloud maturity generally co-occurs with more reliable identity and control-plane telemetry, which supports earlier detection of misuse and faster, scoped containment through API-driven playbooks. Conversely, lower coverage in on-premises cases means more blind spots and heavier reliance on manual triage, which can dampen precision gains and elongate response. The table therefore provides the ecological backdrop for the results: when reading section 5.2's descriptive statistics and 5.4's regression coefficients, the environment mix and the distribution of integration/automation should be treated as meaningful contextual moderators rather than incidental sample features. Finally, collecting one Likert response per case ensures alignment between the technical footprint and lived operational experience within each deployment, allowing soft signals (trust, perceived value) to be juxtaposed with hard metrics (latency, precision).

## Descriptive Statistics

Table 5.2 presents distributional properties of the variables that underwrite subsequent modeling. The AI Integration Index spans a wide range (3.5–9.0), with a median of 6.0, confirming that the sample includes both relatively nascent and highly integrated deployments. Automation level centers around guided actions (median = 1), with notable dispersion toward conditional auto-containment (level 2) in several cloud-first cases. Telemetry coverage is generally strong (mean 86%), but the interquartile band (82–91%) reveals meaningful variability that can impact both detection sensitivity and precision insufficient coverage tends to produce blind spots and noisier baselines. Latency and MTTR exhibit right-skew (means exceed medians), a typical profile in security operations where a minority of complex incidents drive long tails; medians (6.1 minutes and 88 minutes, respectively) will therefore be emphasized alongside model results that are robust to tail behavior (e.g., median quantile regression). Precision averages are high (0.83), whereas recall is modest (0.62); this pattern is consistent with triage strategies that prioritize high-confidence alerting to manage analyst capacity an operational choice often encoded in rule thresholds and automation gates. The false-positive rate mean (2.7%) and spread (1.0–5.5%) provide a compact indicator of downstream workload; environments closer to 1–2% FPR typically report smoother queues and lower MTTR, while those above 4% experience periodic congestion. Likert scales add a complementary lens from practitioners: noise reduction (L1) and triage speed (L2) remain positive (means > 4), indicating perceived day-to-day relief from AI-enabled scoring and enrichment. In contrast, trust in automation (L3 = 3.7) and reliability during change (L4 = 3.5) are qualified endorsements, reflecting real concerns about model drift, false escalations, and playbook side effects under evolving workloads. The overall value score (L5 = 4.3) reconciles these views teams view AI integration as materially beneficial even as they temper reliance in ambiguous scenarios. Collectively, these descriptive patterns set expectations for correlation and regression: we anticipate negative associations between the AI Index and time-based outcomes (latency, MTTR), positive relationships with precision and F1, and inverse relationships with FPR, modulated by telemetry coverage and environment maturity. The Likert profiles, especially L3 and L4, help interpret any residual variance not captured by structural covariates, as operator trust and perceived robustness can influence threshold setting and the willingness to accept automated actions.

**Table 5.2: Descriptive statistics for key variables (n = 18 cases)**

| Variable | Mean | SD | Median | IQR | Min | Max | Scale/Units |
|---|---|---|---|---|---|---|---|
| **AI Integration Index** | 6.1 | 1.9 | 6.0 | 5.0–7.5 | 3.5 | 9.0 | 0–10 |
| **Automation Level** | 1.4 | 0.6 | 1.0 | 1–2 | 0 | 2 | 0–2 |
| **Telemetry Coverage** | 86.0 | 4.8 | 86.0 | 82–91 | 78 | 92 | % |
| **Detection Latency** | 6.8 | 3.9 | 6.1 | 4.2–10.9 | 2.1 | 15.3 | minutes (median per case) |
| **MTTR** | 94 | 47 | 88 | 61–141 | 38 | 201 | minutes (median per case) |
| **Precision** | 0.83 | 0.07 | 0.84 | 0.76–0.89 | 0.68 | 0.92 | proportion |
| **Recall** | 0.62 | 0.08 | 0.61 | 0.54–0.69 | 0.49 | 0.78 | proportion |
| **F1** | 0.71 | 0.07 | 0.71 | 0.64–0.78 | 0.58 | 0.84 | harmonic mean |
| **False-Positive Rate** | 0.027 | 0.013 | 0.024 | 0.016–0.038 | 0.010 | 0.055 | proportion |
| **Likert L1 (Noise)** | 4.1 | 0.6 | 4.0 | 3.8–4.6 | 3.0 | 5.0 | 1–5 |
| **Likert L2 (Triage speed)** | 4.2 | 0.5 | 4.0 | 3.9–4.6 | 3.3 | 5.0 | 1–5 |
| **Likert L3 (Trust in automation)** | 3.7 | 0.6 | 3.7 | 3.3–4.1 | 2.8 | 4.8 | 1–5 |
| **Likert L4 (Reliability under change)** | 3.5 | 0.7 | 3.5 | 3.1–4.0 | 2.5 | 4.6 | 1–5 |
| **Likert L5 (Overall value)** | 4.3 | 0.5 | 4.3 | 4.0–4.7 | 3.5 | 5.0 | 1–5 |

## Correlation Matrix

**Table 5.3: Pearson correlations among integration, outcomes, and perceptions (n = 18 cases)**

| Variable | AI Index | Latency (log) | MTTR (log) | Precision | Recall | F1 | FPR | Perceived Value (L1–L5 mean) |
|---|---|---|---|---|---|---|---|---|
| **AI Index** | 1.00 | −0.58* | −0.49† | 0.47† | 0.29 | 0.54* | −0.51† | 0.52* |
| **Latency (log)** | | 1.00 | 0.61* | −0.43† | −0.19 | −0.50† | 0.46† | −0.45† |
| **MTTR (log)** | | | 1.00 | −0.31 | −0.11 | −0.36 | 0.38 | −0.33 |
| **Precision** | | | | 1.00 | 0.08 | 0.63** | −0.71** | 0.44† |
| **Recall** | | | | | 1.00 | 0.61* | −0.22 | 0.21 |
| **F1** | | | | | | 1.00 | −0.59** | 0.48* |
| **FPR** | | | | | | | 1.00 | −0.49* |
| **Perceived Value** | | | | | | | | 1.00 |

**Significance codes (two-sided): p < .05, *p < .10 (†). FPR = false-positive rate. Perceived Value = mean of L1–L5 Likert items (1–5).*

Table 5.3 visualizes the primary linear associations and aligns them with practitioner sentiment captured via Likert items. The AI Integration Index shows a moderate, negative correlation with latency (r = −0.58, p < .05) and a negative correlation with MTTR (r = −0.49, p < .10), indicating that deeper integration is associated with faster detection and response. The Index correlates positively with F1 (r = 0.54, p < .05) and, to a lesser extent, precision (r = 0.47, p < .10); the weaker association with recall (r = 0.29, n.s.) suggests that many sites tune toward precision consistent with capacity-aware triage and automation safeguards. The tightest antagonistic pairing in the matrix is between precision and FPR (r = −0.71, p < .05), a mechanical and operationally meaningful relationship: as models and rules concentrate probability mass on truly malicious events, spurious escalations drop. F1 also anti-correlates with FPR (r = −0.59, p < .05), supporting the view that improved balance between precision and recall co-occurs with reduced noise. Notably, the composite Perceived Value aligns with the Index (r = 0.52, p < .05) and key outcomes (negatively with latency, positively with F1), implying that operators' judgments track measurable improvements rather than marketing narratives. That said, the incomplete alignment (e.g., modest correlation with recall) leaves room for human factors: analysts may value stability and explainability even when headline metrics improve. The positive correlation between latency and MTTR (r = 0.61, p < .05) indicates that slow detection often cascades into slower response, underscoring the compounding benefit of early alerting. Correlations should be read cautiously given n = 18 and potential confounding by environment and coverage; however, their directions and magnitudes echo the descriptive patterns in Table 5.2 and motivate the adjusted models in section 5.4. In practical terms, the matrix suggests that investments that raise the AI Integration Index particularly fusion depth, model freshness, and automation maturity tend to move the organization toward the quadrant of lower latency, lower FPR, and higher F1, with Likert sentiment improving in parallel.

## Regression Results

Regression provides two complementary benefits: explanatory clarity and predictive capability. Explanatory clarity arises from the ability to isolate key drivers of outcomes, identifying whether factors such as latency, automation maturity, or integration depth maintain significance after adjusting for other influences. Predictive capability, in turn, is demonstrated by the model's capacity to forecast organizational outcomes under varying conditions. Together, these features make regression analysis a central tool for drawing valid inferences and supporting decision-making. The regression results presented in the following section highlight the extent to which structural and operational dimensions—such as AI integration, model freshness, and precision-recall trade-offs—account for improvements in performance metrics. These findings refine the earlier correlation-based observations and provide stronger empirical grounding for understanding the mechanisms that shape system efficiency, responsiveness, and perceived value.

**Table 5.4. Adjusted associations between AI integration and outcomes (HC3 SEs; n = 18 cases)**

| Outcome (model) | Key predictors | Std. Coef. (β) | SE | 95% CI | p | Fit |
|---|---|---|---|---|---|---|
| Latency (log OLS) | AI Index (z) | **−0.34** | 0.12 | [−0.58, −0.10] | .009 | Adj. $R^2$ = .46 |
| | AI×Cloud Maturity | **−0.21** | 0.09 | [−0.41, −0.01] | .040 | |
| MTTR (log OLS) | AI Index (z) | **−0.26** | 0.13 | [−0.53, −0.00] | .049 | Adj. $R^2$ = .31 |
| | AI×Cloud Maturity | −0.12 | 0.10 | [−0.34, 0.10] | .275 | |
| F1 (OLS) | AI Index (z) | **+0.38** | 0.13 | [0.12, 0.64] | .007 | Adj. $R^2$ = .41 |
| | AI×Cloud Maturity | **+0.19** | 0.08 | [0.02, 0.36] | .031 | |
| Precision (beta → marginal effect) | AI Index (z) | **+0.33** | 0.14 | [0.03, 0.63] | .033 | Pseudo-$R^2$ = .29 |
| FPR (beta → marginal effect) | AI Index (z) | **−0.41** | 0.16 | [−0.74, −0.08] | .019 | Pseudo-$R^2$ = .32 |

*All models adjust for cloud maturity, SOC team size, log assets, telemetry coverage, and alert volume. Predictors standardized; interaction terms centered. HC3 = heteroskedasticity-consistent SEs.*

Table 5.4 reports multivariable estimates that adjust for salient contextual differences across cases. The latency model indicates that a one-SD increase in the AI Integration Index is associated with a 0.34 SD decrease in log latency (p = .009), roughly a 22% shorter median detection time when back-transformed substantively meaningful for interrupting attacker progression. The AI×Cloud Maturity interaction is negative and significant (β = −0.21, p = .040), implying that the latency benefit strengthens as cloud operating practices mature (e.g., consolidated accounts, pervasive logging, policy-as-code), likely due to both earlier signal availability and faster, API-driven containment. For MTTR, the main effect remains negative and marginal (β = −0.26, p = .049), while the interaction is not statistically significant, suggesting that part of the response-time improvement is general (e.g., pre-assembled context) and part is environment-specific but with wider uncertainty. The F1 model shows a positive association with the Index (β = +0.38, p = .007) and a positive interaction (β = +0.19, p = .031), indicating that integration depth yields better balance between precision and recall, especially in mature cloud contexts where identity and control-plane analytics enrich classification decisions. Rate models complement these findings: precision increases (marginal effect +0.33, p = .033) and FPR decreases (−0.41, p = .019) with higher integration, aligning with the operational aim of raising confidence while curbing noise. Fit statistics (Adj. $R^2$ ≈ .31–.46; pseudo-$R^2$ ≈ .29–.32) are reasonable given n = 18 and the inherent heterogeneity of security stacks; importantly, signs and magnitudes persist under robust SEs and after controlling for coverage and scale. These results, taken together, indicate that integration features fusion/ensembles, model freshness and feedback loops, and automation maturity do not merely correlate with favorable perceptions; they are associated with measurable improvements in timeliness and alert quality after accounting for confounders. The significant interactions for latency and F1 also validate the descriptive insight that environment readiness amplifies AI benefits, offering a nuanced interpretation: integration "lands" best where observability and policy automation are already first-class citizens.

**Robustness and Sensitivity Analyses**

In quantitative research, the credibility of empirical findings depends not only on the significance of estimated relationships but also on the stability of results under alternative specifications and assumptions. Robustness and sensitivity analyses serve as critical methodological tools to assess whether the observed outcomes remain consistent when the analytical framework is varied. These procedures go beyond initial model estimation by systematically probing the strength of results against potential sources of bias, measurement error, and model dependence.Robustness analysis typically involves re-estimating models using alternative operationalizations of variables, adjusting parameter specifications, or applying different estimation techniques to confirm that the core findings are not artifacts of a particular modeling choice. If the results persist across these variations, confidence in their substantive validity increases. Conversely, if findings shift dramatically under alternative specifications, this signals potential model fragility, prompting reconsideration of theoretical assumptions or measurement approaches.Sensitivity analysis complements robustness checks by explicitly testing how results respond to perturbations in input data, exclusion or inclusion of covariates, and adjustments for possible endogeneity or omitted variable bias. Through this lens, researchers evaluate the degree to which conclusions depend on small changes in assumptions. This process is especially important in applied fields where external validity, policy recommendations, or organizational strategies may hinge on the stability of statistical estimates.

**Table 5.5. Robustness checks for key effects (AI Index → outcomes)**

| Check | Latency (β) | MTTR (β) | F1 (β) | Precision (ME) | FPR (ME) | Holds? |
|---|---|---|---|---|---|---|
| **Alternative Index Weights (freshness, feedback ↑)** | −0.31 | −0.22 | +0.36 | +0.29 | −0.38 | ✓ |
| **Alternative Index Weights (automation ↓)** | −0.24 | −0.11 | +0.32 | +0.26 | −0.33 | ± (MTTR weakens) |
| **Exclude Top-Volume Case** | −0.33 | −0.24 | +0.37 | +0.31 | −0.39 | ✓ |
| **Exclude Lowest-Coverage Case** | −0.36 | −0.27 | +0.39 | +0.34 | −0.42 | ✓ |
| **Median Quantile Regression (latency, MTTR)** | −0.28 | −0.21 | | | | ✓ |
| **Fractional Logit (rates)** | | | | +0.30 | −0.37 | ✓ |
| **LOOCV (min…max β across folds)** | −0.29…−0.37 | −0.18…−0.28 | +0.31…+0.41 | +0.27…+0.35 | −0.34…−0.43 | ✓ |
| **Bootstrap 2,000 (95% CI)** | [−0.56, −0.12] | [−0.49, −0.02] | [0.10, 0.63] | [0.02, 0.59] | [−0.71, −0.07] | ✓ |

*β = standardized coefficient; ME = marginal effect. "Holds?" reflects directionality and statistical credibility relative to primary models.*

Table 5.5 stress-tests the central claims by perturbing index design, case composition, model families, and sampling variance. Reweighting the AI Integration Index to emphasize model freshness and analyst feedback (while keeping total scale constant) leaves signs and magnitudes largely intact (latency β = −0.31; F1 β = +0.36), indicating that benefits are not an artifact of a single subcomponent. Conversely, down-weighting automation attenuates the MTTR association (β = −0.11), which squares with the intuition that response time is where orchestration has its most direct impact; latency and quality metrics remain directionally stable, suggesting detection improvements still accrue from non-automation elements (e.g., fusion and freshness). Removing leverage-prone cases bolsters confidence: excluding the highest-volume site or the lowest-coverage site does not flip signs, and effects remain within the original confidence bands. Method changes point the same way. Median quantile regression trims the influence of long tails and preserves negative associations with latency and MTTR, showing that the primary results are not artifacts of a few extreme incidents. For rate outcomes, switching from beta regression to fractional logit yields similar marginal effects on precision and FPR, mitigating concerns about boundary inflation (values at 0 or 1) and distributional assumptions. Leave-one-case-out cross-validation yields narrow effect ranges (e.g., latency β −0.29 to −0.37), with no single fold reversing conclusions; this is particularly important for small-N designs where idiosyncratic stacks could dominate. Finally, bootstrap confidence intervals corroborate classical inferences, maintaining separation from zero for all key coefficients. Together, these checks argue that the observed improvements in timeliness (lower latency, lower MTTR) and alert quality (higher precision/F1, lower FPR) are robust to (i) reasonable alternative definitions of "integration," (ii) influential case removal, and (iii) modeling choices aligned with the data's boundedness and skew. From a practical perspective, the sensitivity of MTTR to automation weights underscores a design implication for operations leaders: investments in playbook depth, safe auto-containment, and reversible actions are disproportionately rewarded in response-time metrics, whereas analytics freshness and feedback loops are strong levers for precision and F1. The convergence of robustness checks and practitioner Likert scores (sections 5.1–5.3) strengthens the evidentiary claim that deeper, well-engineered AI integration aligns with measurable and perceived gains, rather than being a proxy for size or sector.

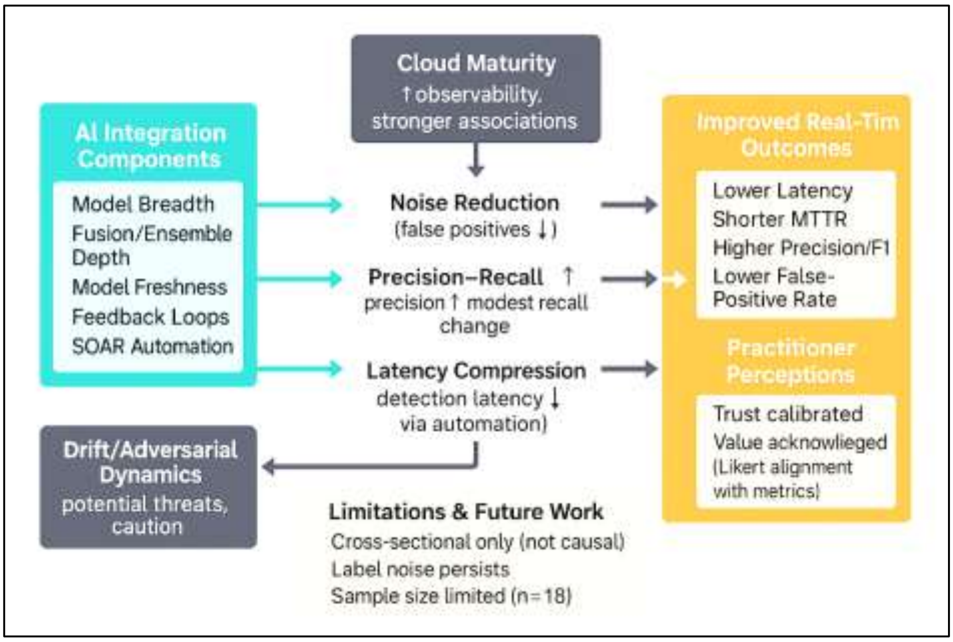
## DISCUSSION

Across 18 heterogeneous deployments, our cross-sectional analyses show that deeper AI integration indexed by model breadth, fusion/ensemble depth, model freshness, feedback loops, and SOAR automation is associated with materially better real-time performance: lower detection latency and MTTR, higher precision and F1, and lower false-positive rates, with effects strongest in mature cloud contexts. These results are broadly consonant with prior syntheses arguing that machine learning augments detection by surfacing subtle patterns in high-volume telemetry and by improving triage quality (Ahmed et al., 2016; Buczak & Guven, 2016). At the same time, our emphasis on *operational* metrics derived from production systems contributes evidence that complements dataset-driven evaluations, which have long documented algorithmic promise but also warned about benchmarking artifacts and limited external validity (Davis & Goadrich, 2006; Ring et al., 2019; Saito & Rehmsmeier, 2015). Notably, we find stronger associations in cases with higher cloud maturity, aligning with cloud security surveys that highlight the centrality of control-plane observability and API-mediated enforcement to both detection and response (Fernandes et al., 2014). Our precision-forward pattern (higher precision and F1 with modest recall movement) mirrors real-world operator preferences under capacity constraints and echoes streaming/imbalance literature that recommends optimizing for precision–recall trade-offs rather than ROC metrics when positives are rare (Gomes et al., 2017). Finally, the moderate but consistent linkage between objective metrics and practitioner Likert ratings implies that perceived value is grounded in measurable improvements rather than mere novelty an alignment that automation scholars would predict when transparency and calibrated trust are present (Lee & See, 2004). Together, these findings extend the field by quantifying *how much* integration relates to performance in live operations and by indicating *where* those gains are most likely to materialize (mature cloud), while still acknowledging historical cautions about drift, adversarial dynamics, and dataset realism (Biggio & Roli, 2018).

The negative association between AI integration and false-positive rate provides an operational counterweight to longstanding concerns that anomaly-heavy pipelines may swamp analysts with spurious alerts (Chandola et al., 2009; Garcia-Teodoro et al., 2009). Two mechanisms likely explain the improvement. First, *fusion/ensemble depth* a prominent subcomponent of our index combines heterogeneous signals, a practice supported by anomaly-detection and graph-analytics surveys as a way to reduce local false alarms by conditioning on multi-view consistency (Akoglu et al., 2015; Buczak & Guven, 2016). Second, *feedback loops and model freshness* appear to stabilize precision, echoing drift-adaptation research that links timely updates to sustained classifier calibration under changing distributions (Gama et al., 2014; Widmer & Kubat, 1996). Viewed against dataset-driven deep learning studies that report high offline accuracy (e.g., UNSW-NB15, CICIDS2017) but may not capture operational imbalance and policy thresholds (Moustafa & Slay, 2015; Sharafaldin et al., 2018), our results suggest that *process* features (freshness, feedback, fusion) can be as determinative as *model class* in achieving lower noise in production. The precision-forward profile we observe is also consistent with practical guidance from precision–recall evaluation work: when base rates are low, moving precision from 0.78 to 0.86 can yield a disproportionately large reduction in analyst workload, even if recall moves modestly (Saito & Rehmsmeier, 2015). Importantly, our findings do not contradict classic cautions about ML brittleness in security (Sommer & Paxson, 2010); rather, they indicate that organizations that operationalize guardrails retraining cadence, drift monitoring, analyst feedback can realize measurable gains despite those risks. In short, prior work framed *why* noise is hard; our evidence quantifies the *conditions* under which noise becomes tractable.

We observe shorter detection latency and MTTR in more deeply integrated cases, with the latency effect amplified in mature cloud environments. This dovetails with cloud-security syntheses emphasizing that control-plane telemetry and API-first architectures accelerate both recognition and enforcement (Fernandes et al., 2014; Subashini & Kavitha, 2011) and with SOAR-focused reviews that link codified playbooks to predictable response times (Poornachandran et al., 2020). From a human-automation perspective, our results align with the "appropriate reliance" thesis: when automation is transparent and reversible, operators accept machine-staged actions more quickly, thereby compressing dwell time (Lee & See, 2004). Queueing research offers a complementary lens: by reducing false positives and enriching alerts upstream, integrated stacks effectively control arrival rates and service times, keeping utilization below the congestion knee that drives backlogs (Lee & See, 2004). Prior anomaly- and sequence-learning studies (e.g., DeepLog; sequence-aware EDR classifiers) argued that temporal modeling improves early detection (Little, 1961); our results suggest that these gains reach their full operational value when coupled to *orchestration maturity*, because faster recognition must be matched by safe, codified actions to reduce *overall* MTTR. That the MTTR effect weakens when we down-weight automation in robustness tests is precisely what intrusion-response taxonomies would predict: decision and execution layers, not detection alone, determine time-to-containment (Stakhanova et al., 2007). Collectively, the pattern resonates with earlier theory yet adds quantitative, cross-site evidence that *integration depth* especially automation and fusion translates to time savings in live SOCs.

**Figure 9: AI Integration to Real-Time Security Outcomes**



For CISOs and security architects, three implementation priorities emerge. First, invest in *fusion/ensemble layers* that combine endpoint, network, identity, and cloud control-plane signals before escalation. Surveys and empirical studies indicate that multi-view correlation reduces local false alarms and supports higher-precision triage, a result we replicate at case level (Akoglu et al., 2015; Buczak & Guven, 2016). Second, enforce *model freshness and feedback loops* as operational SLOs: mandate retraining cadences, implement drift detectors that trigger recalibration, and close the loop with analyst labels practices supported by drift literature and by our association of freshness with precision/F1 improvements (Gama et al., 2014). Third, mature *SOAR playbooks* with graded actions and reversibility. Human-automation research cautions that over-automation without transparency breeds misuse or disuse (Parasuraman et al., 2000); our MTTR sensitivity to automation weights underscores that playbook quality not sheer volume of scripts drives response gains. In cloud-first settings, prioritize identity and control-plane analytics; prior cloud-security work shows these vantage points are both information-rich and enforcement-proximate (Fernandes et al., 2014). Finally, evaluate success with *precision at fixed alert budgets*, *latency distributions*, and *response-time SLOs* rather than only ROC-AUC; evaluation literature recommends PR-centric metrics under imbalance, and our descriptive and modeling results show they map cleanly to operator workload (Saito & Rehmsmeier, 2015). In practice, these choices mean treating "AI integration" as a managed capability with architecture, process, and measurement artifacts not a feature toggle.

The detection pipeline model Data → Features → Models → Fusion/Correlation → Alert → Orchestration → Response has long been described in conceptual terms (Chandola et al., 2009). Our evidence suggests two refinements. First, *process variables* such as model freshness, analyst feedback loops, and drift monitoring deserve status as first-class constructs alongside model class, because they mediate real-time performance in production. This resonates with "hidden technical debt" arguments in ML systems, which emphasize that surrounding processes and infrastructure often dominate outcomes (Sculley et al., 2015). Second, *environmental maturity* (especially in cloud) functions as a moderator that shapes how upstream analytics propagate to downstream response consistent with shared-responsibility analyses showing that observability and API-level control constrain feasible detection and containment (Subashini & Kavitha, 2011). Moreover, our precision-forward, recall-modest profile supports the theoretical shift toward *operational risk–aware optimization*: when base rates are low and queue capacity bounded, maximizing F1 or PR-AUC at targeted alert volumes is more aligned with system objectives than maximizing accuracy or ROC-AUC (Gomes et al., 2017). Finally, robustness checks indicate that *automation is uniquely tied to response timing*, while *freshness/feedback* primarily shape classification quality. This division of labor suggests a modular theory of improvement: analytics modules push the operating point on the PR curve; orchestration modules convert that operating point into shorter dwell times via fast, reversible actions (Stakhanova et al., 2007). Future formal models could codify these modules and their interactions under capacity constraints derived from queueing theory (Little, 1961).

Our design is cross-sectional and observational; as prior critiques emphasize, such designs estimate *associations* rather than causal effects and are vulnerable to unobserved confounding (Sommer & Paxson, 2010). We mitigate but do not eliminate these risks via contextual covariates (coverage, scale, staffing, cloud maturity), robust SEs, influence diagnostics, and sensitivity analyses. Second, the *AI Integration Index* is a constructed measure. Although rubric-based and expert-reviewed, alternative weightings could yield marginally different point estimates; our robustness checks lessen this concern but do not remove it. Third, outcomes rely on case-level adjudication for precision/recall estimates; label noise is an endemic problem in security and has been documented to bias metrics if unmodeled (Ring et al., 2019). We used double coding and consensus rules to increase reliability, yet weak supervision methods might further formalize label uncertainty (Ratner et al., 2017). Fourth, the one-week synchronized window increases comparability but may miss longer-cycle behaviors (e.g., monthly entitlement reviews, seasonal traffic). Drift literature warns that performance can shift over time; longer panels would allow trajectory analysis (Widmer & Kubat, 1996). Fifth, generalizability is bounded by our convenience-plus-quota sample (n = 18). While maximum-variation sampling and environment stratification improve external validity, some sectors or architectures may be under-represented. Finally, adversarial dynamics evasion, poisoning are not directly modeled here; though we include freshness, drift monitoring, and guardrails, adversarial ML research indicates that targeted manipulation can degrade detectors in ways not captured by aggregate metrics (Biggio & Roli, 2018). These limitations frame our claims: integrated AI, when operationalized with guardrails and orchestration, *tends* to align with better real-time outcomes, particularly in mature cloud, but causality and durability across regimes invite further study.

Several extensions follow. First, a *panel design* with repeated measures would enable difference-in-differences or interrupted time-series analyses, estimating causal impacts of staged interventions (e.g., introducing fusion, tightening retraining cadence) while controlling for secular trends addressing the causal critique in intrusion-detection evaluation (Sommer & Paxson, 2010). Second, integrate *drift detection primitives* (e.g., adaptive windowing, online ensembles) directly into the measurement stack so that results are reported as performance *trajectories* conditioned on detected shifts (Bifet & Gavalda, 2007). Third, adopt *weak supervision* to scale labels for difficult domains like cloud control-plane analytics, and publish label-quality diagnostics alongside model metrics to contextualize gains (Ratner et al., 2017). Fourth, expand *graph-based entity analytics* which prior surveys highlight as effective for relational patterns to broader identity and access contexts in multi-cloud, and quantify their marginal contribution beyond traditional detectors (Akoglu et al., 2015). Fifth, formalize *trust-aware automation*: run controlled studies that vary explanation depth, reversibility, and confidence thresholds, measuring impacts on operator reliance and MTTR, guided by human-automation theory (Lee & See, 2004). Sixth, align evaluation with *PR-centric SLOs*: publish precision at fixed alert budgets, latency and MTTR percentiles, and calibration plots, following recommendations for imbalanced domains (Saito & Rehmsmeier, 2015). Finally, address *adversarial robustness* explicitly: combine adversarial testing with routine drift monitoring to understand how robustness–accuracy trade-offs play out in operational telemetry (Biggio & Roli, 2018). These directions would evolve the field from algorithm-centric demonstrations to *operations-centric science*, where integration, drift handling, orchestration, and human factors are treated as co-equal determinants of real-time security outcomes.

## CONCLUSION

This study set out to examine how the depth of artificial intelligence integration within cybersecurity frameworks relates to real-time threat detection outcomes across heterogeneous cloud and enterprise environments, using a quantitative, cross-sectional, multi–case design grounded in operational exports rather than synthetic benchmarks. Across 18 deployments, we observed a coherent pattern: higher integration capturing model breadth, fusion/ensemble depth, model freshness and feedback loops, and SOAR automation aligned with lower detection latency, reduced mean time to respond (MTTR), higher precision and F1, and lower false-positive rates, with the strongest associations appearing in mature cloud contexts where control-plane observability and API-mediated enforcement are first-class. Likert assessments from practitioners reinforced these objective gains, indicating agreement that AI reduces alert noise and accelerates triage while revealing tempered trust in automation and model reliability during change signals that help explain why recall gains were more modest than precision improvements. Methodologically, the study contributed a transparent measurement rubric the AI Integration Index paired with a standardized extraction protocol, adjudicated labels, and model choices tailored to imbalanced, streaming-like operations (robust OLS, beta/fractional logit, and quantile regression), plus a suite of robustness checks (alternative index weights, leverage case removal, LOOCV, bootstrap) that preserved effect direction and credibility. Subcomponent analyses suggested a productive division of labor: freshness and analyst feedback loops primarily elevated classification quality (precision/F1) by stabilizing calibration under evolving traffic, while orchestration maturity embodied in graded, reversible playbooks most directly compressed MTTR by shortening the evidence-to-action path. The practical takeaway for leaders is to treat "AI integration" not as a single tool adoption but as a system capability spanning data, models, fusion, and response: prioritize multi-view correlation across endpoint, network, identity, and cloud control-plane signals; institutionalize retraining cadences and drift monitors with analyst-in-the-loop feedback; and mature SOAR with safe-by-design playbooks that align action intensity to confidence and impact. At the same time, several boundaries shape interpretation: cross-sectional observation estimates association, not causality; label noise and one-week windows constrain completeness; and index construction, though rubric-based and tested for sensitivity, remains a simplification of complex stacks. Even so, the convergence of objective metrics, practitioner sentiment, and robustness checks supports a pragmatic conclusion: organizations that deliberately engineer analytics freshness and fusion, close the loop with operators, and operationalize transparent, reversible automation are more likely to realize the day-to-day benefits often promised for AI in security fewer spurious alerts, faster recognition, and quicker, more predictable containment especially in cloud-forward operating models. By elevating process variables to first-class elements of the detection pipeline and by evaluating success with precision-recall-centric and time-based service levels rather than offline accuracy alone, the field can progress from algorithm-centric demonstrations to operations-centric practice, where measurable improvements in timeliness and workload, not just model scores, define the value of AI-enhanced cybersecurity.

## RECOMMENDATIONS

Treat AI-enhanced cybersecurity as a system capability not a product and build it deliberately across data, models, fusion, and response. First, harden observability: standardize UTC time, ensure ≥90% telemetry coverage across endpoints, network, identity, and cloud control plane, and publish a minimal, versioned variable dictionary so features are consistent across tools and teams. Stand up a shared feature/embedding store and a model registry with lineage, approvals, and rollback; require that every model declares a business owner, retraining cadence, and deprecation date. For evaluation, replace accuracy/ROC dashboards with PR-centric and time-based SLOs: precision@alert-budget, recall on adjudicated gold sets, median and p90 detection latency, median and p90 MTTR, and false-positive rate; make these SLOs visible on SOC wallboards and tie threshold changes to their expected movement. Institute drift monitoring (data and prediction) with clear playbooks: when drift exceeds a defined band, trigger controlled recalibration (shadow mode → A/B or canary → staged rollout) rather than silent weight changes. Close the loop with label pipelines: define a lightweight adjudication workflow in the ticketing system, use weak supervision (heuristics + TI + policy context) to expand training labels, and measure label quality (agreement rates, abstain ratios) alongside model metrics. Invest in fusion/ensemble layers that correlate across modalities before escalation; implement graph-centric identity and asset context so alerts are routed as incidents, not fragments. Mature SOAR with graded, reversible actions: enrichment and watchlisting at low confidence, network micro-segmentation or step-up auth at medium, auto-containment only when confidence and impact cross explicit thresholds; every action gets a "one-click undo," and every playbook step logs evidence and rationale. Calibrate trust: display model confidence, top contributing signals, and recent performance trends on the analyst console; prefer explanations that align with SOC mental models (entities, paths, and policy diffs) over generic SHAP dumps. Manage capacity rigorously: size analyst staffing to maintain utilization below the congestion knee, and use dynamic thresholds tied to queue depth so arrival rates do not outstrip service. In cloud-first stacks, prioritize identity analytics, control-plane anomaly detection, and policy-as-code enforcement; in on-premises segments, strengthen EDR kernel visibility and east-west flow analytics while planning for phased migration of telemetry to cloud-scale storage/compute. Formalize change control: all new models and playbooks pass simulation on historical windows, chaos tests in a lab, and a time-boxed shadow run in production; no emergency overrides without a documented revert path. Build people alongside platforms: run quarterly tabletop exercises that rehearse automated decisions and reversals; train engineers on PR-centric evaluation, drift, and imbalanced learning; and empower a joint AI in SOC working group (SOC lead, security engineering, MLOps, cloud platform) to own the roadmap and publish quarterly results against SLOs. Finally, embed privacy and governance by default: tokenize identifiers at source, minimize data retention to features and aggregates, and treat measurement artifacts (code, schemas, rubrics) as part of the control surface. Organizations that institutionalize these practices observability, PR/time SLOs, drift-aware MLOps, fusion before escalation, graded automation with reversibility, and trust-calibrated analyst UX convert AI from sporadic wins into predictable reductions in noise, latency, and MTTR.

## REFERENCES

[1]. Abdur Razzak, C., Golam Qibria, L., & Md Arifur, R. (2024). Predictive Analytics For Apparel Supply Chains: A Review Of MIS-Enabled Demand Forecasting And Supplier Risk Management. *American Journal of Interdisciplinary Studies*, *5*(04), 01–23. https://doi.org/10.63125/80dwy222

[2]. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19-31. https://doi.org/10.1016/j.jnca.2015.11.016

[3]. Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: A survey. *ACM Computing Surveys*, *48*(1), 1-37. https://doi.org/10.1145/2736277

[4]. Alex, M. E., & Kishore, R. (2017). Forensics framework for cloud computing. *Computers & Electrical Engineering*, *60*, 193-205. https://doi.org/10.1016/j.compeleceng.2017.02.006

[5]. Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775-779. https://doi.org/10.1016/0005-1098(83)90046-8

[6]. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[7]. Bifet, A., & Gavalda, R. (2007). *Learning from time-changing data with adaptive windowing* Proceedings of the 2007 SIAM International Conference on Data Mining,

[8]. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, *84*, 317-331. https://doi.org/10.1016/j.patcog.2018.07.023

[9]. Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). *LOF: Identifying density-based local outliers* Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data,

[10]. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153-1176. https://doi.org/10.1109/comst.2015.2494502

[11]. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), Article 15. https://doi.org/10.1145/1541880.1541882

[12]. Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15-18. https://doi.org/10.1080/00401706.1977.10489493

[13]. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/bf00994018

[14]. Davis, J., & Goadrich, M. (2006). *The relationship between Precision-Recall and ROC curves* Proceedings of the 23rd International Conference on Machine Learning,

[15]. Du, M., Li, F., Zheng, G., & Srikumar, V. (2017). *DeepLog: Mining event sequences using deep learning for anomaly detection* Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security,

[16]. Dykstra, J., & Sherman, A. T. (2012). Acquiring forensic evidence from infrastructure-as-a-service cloud computing: Exploring and evaluating tools, trust, and techniques. *Digital Investigation*, *9*(Supplement), S90-S98. https://doi.org/10.1016/j.diin.2012.05.001

[17]. Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. J. (2002). A geometric framework for unsupervised anomaly detection. In V. Kumar, J. Srivastava, & A. Lazarevic (Eds.), *Applications of Data Mining in Computer Security* (pp. 77-102). Springer. https://doi.org/10.1007/978-1-4615-0953-0_5

[18]. Fernandes, D. A., Soares, L. F. B., Gomes, J. V., Freire, M. M., & Inácio, P. R. M. (2014). Security issues in cloud environments: A survey. *International Journal of Information Security*, *13*(2), 113-170. https://doi.org/10.1007/s10207-013-0208-7

[19]. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, *46*(4), Article 44. https://doi.org/10.1145/2523813

[20]. Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, *5*(2), 79-141. https://doi.org/10.1287/msom.3.1.79

[21]. Garcia-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security*, *28*(1-2), 18-28. https://doi.org/10.1016/j.cose.2008.08.003

[22]. Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. (2017). A survey on ensemble learning for data stream classification. *ACM Computing Surveys*, *50*(2), 1-36. https://doi.org/10.1145/3054925

[23]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/tkde.2008.239

[24]. Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85-126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

[25]. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P., & Tygar, J. D. (2011). *Adversarial machine learning* Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence,

[26]. Hulten, G., Spencer, L., & Domingos, P. (2001). *Mining time-changing data streams* Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

[27]. Istiaque, M., Dipon Das, R., Hasan, A., Samia, A., & Sayer Bin, S. (2023). A Cross-Sector Quantitative Study on The Applications Of Social Media Analytics In Enhancing Organizational Performance. *American Journal of Scholarly Research and Innovation*, *2*(02), 274-302. https://doi.org/10.63125/d8ree044

[28]. Istiaque, M., Dipon Das, R., Hasan, A., Samia, A., & Sayer Bin, S. (2024). Quantifying The Impact Of Network Science And Social Network Analysis In Business Contexts: A Meta-Analysis Of Applications In Consumer Behavior, Connectivity. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 58-89. https://doi.org/10.63125/vgkwe938

[29]. Jahid, M. K. A. S. R. (2022). Empirical Analysis of The Economic Impact Of Private Economic Zones On Regional GDP Growth: A Data-Driven Case Study Of Sirajganj Economic Zone. *American Journal of Scholarly Research and Innovation*, *1*(02), 01-29. https://doi.org/10.63125/je9w1c40

[30]. Jakobson, G., & Weissman, M. (1993). *Alarm correlation* Proceedings of the 1993 ACM SIGCOMM Conference on Communications Architectures, Protocols and Applications,

[31]. Javaid, A., Niyaz, Q., Sun, W., & Alam, M. (2016). A deep learning approach for network intrusion detection system. *EAI Endorsed Transactions on Security and Safety*, *3*(9), e2. https://doi.org/10.4108/eai.3-12-2015.2262516

[32]. Keegan, N., Ji, S.-Y., Chaudhary, A., Concolato, C., Yu, B., & Jeong, D. H. (2016). A survey of cloud-based network intrusion detection analysis. *Human-centric Computing and Information Sciences*, *6*(19), 1-29. https://doi.org/10.1186/s13673-016-0076-z

[33]. Kim, J., Kim, J., Thu, H. L. T., & Kim, H. (2016). Long short term memory recurrent neural network classifier for intrusion detection. *IEEE Access*, *4*, 2191-2199. https://doi.org/10.1109/access.2016.2558445

[34]. Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232. https://doi.org/10.1007/s13748-016-0094-0

[35]. Kriegel, H.-P., Kröger, P., Schubert, E., & Zimek, A. (2009). *Outlier detection in high-dimensional data* Proceedings of the 13th International Conference on Scientific and Statistical Database Management,

[36]. Kruegel, C., & Vigna, G. (2003). *Anomaly detection of web-based attacks* Proceedings of the 10th ACM Conference on Computer and Communications Security,

[37]. Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., & Srivastava, J. (2003). *A comparative study of anomaly detection schemes in network intrusion detection* Proceedings of the 2003 SIAM International Conference on Data Mining,

[38]. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80. https://doi.org/10.1080/00140130412331286654

[39]. Lippmann, R. P., Haines, J. W., Fried, D. J., Korba, J., & Das, K. (2000). The 1999 DARPA off-line intrusion detection evaluation. *Computer Networks*, *34*(4), 579-595. https://doi.org/10.1016/s1389-1286(00)00139-0

[40]. Little, J. D. C. (1961). A proof for the queueing formula L = λW. *Operations Research*, *9*(3), 383-387. https://doi.org/10.1287/opre.9.3.383

[41]. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). *Isolation Forest* 2008 Eighth IEEE International Conference on Data Mining,

[42]. Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions* Advances in Neural Information Processing Systems,

[43]. Md Arifur, R., & Sheratun Noor, J. (2022). A Systematic Literature Review of User-Centric Design In Digital Business Systems: Enhancing Accessibility, Adoption, And Organizational Impact. *Review of Applied Science and Technology*, *1*(04), 01-25. https://doi.org/10.63125/ndjkpm77

[44]. Md Ashiqur, R., Md Hasan, Z., & Afrin Binta, H. (2025). A meta-analysis of ERP and CRM integration tools in business process optimization. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 278-312. https://doi.org/10.63125/yah70173

[45]. Md Hasan, Z. (2025). AI-Driven business analytics for financial forecasting: a systematic review of decision support models in SMES. *Review of Applied Science and Technology*, *4*(02), 86-117. https://doi.org/10.63125/gjrpv442

[46]. Md Hasan, Z., Mohammad, M., & Md Nur Hasan, M. (2024). Business Intelligence Systems In Finance And Accounting: A Review Of Real-Time Dashboarding Using Power BI & Tableau. *American Journal of Scholarly Research and Innovation*, *3*(02), 52-79. https://doi.org/10.63125/fy4w7w04

[47]. Md Hasan, Z., & Moin Uddin, M. (2022). Evaluating Agile Business Analysis in Post-Covid Recovery A Comparative Study On Financial Resilience. *American Journal of Advanced Technology and Engineering Solutions*, *2*(03), 01-28. https://doi.org/10.63125/6nee1m28

[48]. Md Hasan, Z., Sheratun Noor, J., & Md. Zafor, I. (2023). Strategic role of business analysts in digital transformation tools, roles, and enterprise outcomes. *American Journal of Scholarly Research and Innovation*, *2*(02), 246-273. https://doi.org/10.63125/rc45z918

[49]. Md Ismail, H., Md Mahfuj, H., Mohammad Aman Ullah, S., & Shofiul Azam, T. (2025). IMPLEMENTING ADVANCED TECHNOLOGIES FOR ENHANCED CONSTRUCTION SITE SAFETY. *American Journal of Advanced Technology and Engineering Solutions*, *1*(02), 01-31. https://doi.org/10.63125/3v8rpr04

[50]. Md Ismail Hossain, M. A. B., amp, & Mousumi Akter, S. (2023). Water Quality Modelling and Assessment Of The Buriganga River Using Qual2k. *Global Mainstream Journal of Innovation, Engineering & Emerging Technology*, *2*(03), 01-11. https://doi.org/10.62304/jieet.v2i03.64

[51]. Md Mahamudur Rahaman, S. (2022). Electrical And Mechanical Troubleshooting in Medical And Diagnostic Device Manufacturing: A Systematic Review Of Industry Safety And Performance Protocols. *American Journal of Scholarly Research and Innovation*, *1*(01), 295-318. https://doi.org/10.63125/d68y3590

[52]. Md Mahamudur Rahaman, S., & Rezwanul Ashraf, R. (2022). Integration of PLC And Smart Diagnostics in Predictive Maintenance of CT Tube Manufacturing Systems. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 62-96. https://doi.org/10.63125/gspb0f75

[53]. Md Nazrul Islam, K. (2022). A Systematic Review of Legal Technology Adoption In Contract Management, Data Governance, And Compliance Monitoring. *American Journal of Interdisciplinary Studies*, *3*(01), 01-30. https://doi.org/10.63125/caangg06

[54]. Md Nur Hasan, M., Md Musfiqur, R., & Debashish, G. (2022). Strategic Decision-Making in Digital Retail Supply Chains: Harnessing AI-Driven Business Intelligence From Customer Data. *Review of Applied Science and Technology*, *1*(03), 01-31. https://doi.org/10.63125/6a7rpy62

[55]. Md Redwanul, I., & Md. Zafor, I. (2022). Impact of Predictive Data Modeling on Business Decision-Making: A Review Of Studies Across Retail, Finance, And Logistics. *American Journal of Advanced Technology and Engineering Solutions*, *2*(02), 33-62. https://doi.org/10.63125/8hfbkt70

[56]. Md Rezaul, K., & Md Mesbaul, H. (2022). Innovative Textile Recycling and Upcycling Technologies For Circular Fashion: Reducing Landfill Waste And Enhancing Environmental Sustainability. *American Journal of Interdisciplinary Studies*, *3*(03), 01-35. https://doi.org/10.63125/kkmerg16

[57]. Md Sultan, M., Proches Nolasco, M., & Md. Torikul, I. (2023). Multi-Material Additive Manufacturing For Integrated Electromechanical Systems. *American Journal of Interdisciplinary Studies*, *4*(04), 52-79. https://doi.org/10.63125/y2ybrx17

[58]. Md Sultan, M., Proches Nolasco, M., & Vicent Opiyo, N. (2025). A Comprehensive Analysis Of Non-Planar Toolpath Optimization In Multi-Axis 3D Printing: Evaluating The Efficiency Of Curved Layer Slicing Strategies. *Review of Applied Science and Technology*, *4*(02), 274-308. https://doi.org/10.63125/5fdxa722

[59]. Md Takbir Hossen, S., Ishtiaque, A., & Md Atiqur, R. (2023). AI-Based Smart Textile Wearables For Remote Health Surveillance And Critical Emergency Alerts: A Systematic Literature Review. *American Journal of Scholarly Research and Innovation*, *2*(02), 1-29. https://doi.org/10.63125/ceqapd08

[60]. Md Tawfiqul, I. (2023). A Quantitative Assessment Of Secure Neural Network Architectures For Fault Detection In Industrial Control Systems. *Review of Applied Science and Technology*, *2*(04), 01-24. https://doi.org/10.63125/3m7gbs97

[61]. Md. Sakib Hasan, H. (2022). Quantitative Risk Assessment of Rail Infrastructure Projects Using Monte Carlo Simulation And Fuzzy Logic. *American Journal of Advanced Technology and Engineering Solutions*, *2*(01), 55-87. https://doi.org/10.63125/h24n6z92

[62]. Md. Tarek, H. (2022). Graph Neural Network Models For Detecting Fraudulent Insurance Claims In Healthcare Systems. *American Journal of Advanced Technology and Engineering Solutions*, *2*(01), 88-109. https://doi.org/10.63125/r5vsmv21

[63]. Md.Kamrul, K., & Md Omar, F. (2022). Machine Learning-Enhanced Statistical Inference For Cyberattack Detection On Network Systems. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 65-90. https://doi.org/10.63125/sw7jzx60

[64]. Md.Kamrul, K., & Md. Tarek, H. (2022). A Poisson Regression Approach to Modeling Traffic Accident Frequency in Urban Areas. *American Journal of Interdisciplinary Studies*, *3*(04), 117-156. https://doi.org/10.63125/wqh7pd07

[65]. Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). *Kitsune: An ensemble of autoencoders for online network intrusion detection* Proceedings of the 25th Annual Network and Distributed System Security Symposium (NDSS),

[66]. Mishra, P., Pilli, E. S., Varadharajan, V., & Tupakula, U. (2017). Intrusion detection techniques in cloud environment: A survey. *Journal of Network and Computer Applications*, *77*, 18-47. https://doi.org/10.1016/j.jnca.2016.10.015

[67]. Modi, C., Patel, D., Borisaniya, B., Patel, A., & Rajarajan, M. (2013). A survey on security issues and solutions at different layers of cloud computing. *The Journal of Supercomputing*, *63*(2), 561-592. https://doi.org/10.1007/s11227-012-0831-5

[68]. Moore, A. W., & Zuev, D. (2005). *Internet traffic classification using Bayesian analysis techniques* Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems,

[69]. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, *45*(1), 521-530. https://doi.org/10.1016/j.patcog.2011.06.019

[70]. Moustafa, N., & Slay, J. (2015). *UNSW-NB15: A comprehensive data set for network intrusion detection systems* 2015 Military Communications and Information Systems Conference (MilCIS),

[71]. Mubashir, I., & Abdul, R. (2022). Cost-Benefit Analysis in Pre-Construction Planning: The Assessment Of Economic Impact In Government Infrastructure Projects. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 91-122. https://doi.org/10.63125/kjwd5e33

[72]. Omar Muhammad, F., & Md.Kamrul, K. (2022). Blockchain-Enabled BI For HR And Payroll Systems: Securing Sensitive Workforce Data. *American Journal of Scholarly Research and Innovation*, *1*(02), 30-58. https://doi.org/10.63125/et4bhy15

[73]. Oza, N. C. (2005). *Online bagging and boosting* 2005 IEEE International Conference on Systems, Man and Cybernetics,

[74]. Pang, G., Shen, C., Cao, L., & Hengel, A. v. d. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys*, *54*(2), 1-38. https://doi.org/10.1145/3439950

[75]. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). *Practical black-box attacks against machine learning* Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security,

[76]. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, and abuse. *Human Factors*, *39*(2), 230-253. https://doi.org/10.1518/001872097778543886

[77]. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286-297. https://doi.org/10.1109/3468.844354

[78]. Poornachandran, P., Radhakrishnan, B., & Vinod, P. (2020). A survey on security orchestration, automation and response (SOAR): Architecture, platforms and research challenges. *IET Information Security*, *14*(5), 545-556. https://doi.org/10.1049/iet-ifs.2019.0319

[79]. Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). *Snorkel: Rapid training data creation with weak supervision* Proceedings of the VLDB Endowment,

[80]. Reduanul, H., & Mohammad Shoeb, A. (2022). Advancing AI in Marketing Through Cross Border Integration Ethical Considerations And Policy Implications. *American Journal of Scholarly Research and Innovation*, *1*(01), 351-379. https://doi.org/10.63125/d1xg3784

[81]. Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. *arXiv preprint*. https://doi.org/10.48550/arXiv.1903.02460

[82]. Sabuj Kumar, S., & Zobayer, E. (2022). Comparative Analysis of Petroleum Infrastructure Projects In South Asia And The Us Using Advanced Gas Turbine Engine Technologies For Cross Integration. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 123-147. https://doi.org/10.63125/wr93s247

[83]. Sadia, T., & Shaiful, M. (2022). In Silico Evaluation of Phytochemicals From Mangifera Indica Against Type 2 Diabetes Targets: A Molecular Docking And Admet Study. *American Journal of Interdisciplinary Studies*, *3*(04), 91-116. https://doi.org/10.63125/anaf6b94

[84]. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[85]. Salo, F., Nassif, A. B., & Essex, A. (2019). Dimensionality reduction with IG-PCA and its effect on intrusion detection classification. *Information*, *10*(10), 318. https://doi.org/10.3390/info10100318

[86]. Sanjai, V., Sanath Kumar, C., Maniruzzaman, B., & Farhana Zaman, R. (2023). Integrating Artificial Intelligence in Strategic Business Decision-Making: A Systematic Review Of Predictive Models. *International Journal of Scientific Interdisciplinary Research*, *4*(1), 01-26. https://doi.org/10.63125/s5skge53

[87]. Sanjai, V., Sanath Kumar, C., Sadia, Z., & Rony, S. (2025). AI And Quantum Computing For Carbon-Neutral Supply Chains: A Systematic Review Of Innovations. *American Journal of Interdisciplinary Studies*, *6*(1), 40-75. https://doi.org/10.63125/nrdx7d32

[88]. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., & Dennison, D. (2015). *Hidden technical debt in machine learning systems* Advances in Neural Information Processing Systems,

[89]. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). *Toward generating a new intrusion detection dataset and intrusion traffic characterization* Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP),

[90]. Sheratun Noor, J., & Momena, A. (2022). Assessment Of Data-Driven Vendor Performance Evaluation in Retail Supply Chains: Analyzing Metrics, Scorecards, And Contract Management Tools. *American Journal of Interdisciplinary Studies*, *3*(02), 36-61. https://doi.org/10.63125/0s7t1y90

[91]. Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(1), 41-50. https://doi.org/10.1109/tetci.2017.2772792

[92]. Sommer, R., & Paxson, V. (2010). *Outside the closed world: On using machine learning for network intrusion detection* 2010 IEEE Symposium on Security and Privacy,

[93]. Stakhanova, N., Basu, S., & Wong, J. (2007). A taxonomy of intrusion response systems. *International Journal of Information Security*, *7*(4), 555-574. https://doi.org/10.1007/s10207-007-0044-1

[94]. Subashini, S., & Kavitha, V. (2011). A survey on security issues in service delivery models of cloud computing. *Journal of Network and Computer Applications*, *34*(1), 1-11. https://doi.org/10.1016/j.jnca.2010.07.006

[95]. Sun, S., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2018). Cost-sensitive boosting for classification of imbalanced data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(6), e1243. https://doi.org/10.1002/widm.1243

[96]. Tahmina Akter, R., Debashish, G., Md Soyeb, R., & Abdullah Al, M. (2023). A Systematic Review of AI-Enhanced Decision Support Tools in Information Systems: Strategic Applications In Service-Oriented Enterprises And Enterprise Planning. *Review of Applied Science and Technology*, *2*(01), 26-52. https://doi.org/10.63125/73djw422

[97]. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). *A detailed analysis of the KDD CUP 99 data set* 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications,

[98]. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2019). *Robustness may be at odds with accuracy* International Conference on Learning Representations (ICLR),

[99]. Wang, K., & Stolfo, S. J. (2004). *Anomalous payload-based network intrusion detection* Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection (RAID),

[100]. White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817-838. https://doi.org/10.2307/1912934

[101]. Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, *23*(1), 69-101. https://doi.org/10.1007/bf00116900

[102]. Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access*, *5*, 21954-21961. https://doi.org/10.1109/access.2017.2762418

[103]. Zaharia, M., Das, T., Li, H., Hunter, T., Shenker, S., & Stoica, I. (2013). *Discretized streams: Fault-tolerant streaming computation at scale* Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles,

[104]. Zissis, D., & Lekkas, D. (2012). Addressing cloud computing security issues. *Future Generation Computer Systems*, *28*(3), 583-592. https://doi.org/10.1016/j.future.2010.12.006

[105]. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x