Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Artificial Intelligence for Cybersecurity and Cloud Security: Predictive Threat Detection, Risk Analytics, and Decision Support for Critical Infrastructure

Shakila Akter[1] and Sajidul Haque Chowdhruy[2]
[1] *Lewis University, Master of Science in Computer and Information Science*
[2] *Washington University of Science and Technology, Master of Science in Cybersecurity*
Corresponding Author: Shakila Akter, **E-mail**: saktershakila23@gmail.com

| ABSTRACT
Artificial intelligence (AI) can strengthen cyber and cloud security by detecting weak signals, estimating operational risk, and prioritizing interventions before disruptions propagate across critical infrastructure. Yet high benchmark accuracy does not by itself establish operational utility: infrastructure operators must control false alarms, account for concept drift, explain recommendations, and preserve human authority over consequential actions. This study develops a governance-aware architecture that fuses network telemetry, identity events, cloud control-plane logs, vulnerability context, asset criticality, and threat intelligence into a calibrated decision-support pipeline. A design-science methodology is combined with a reproducible synthetic benchmark of 60,000 temporally ordered security events, including an induced distribution shift in the final 20 percent. Synthetic data were used because authentic enterprise telemetry is difficult to release without exposing personal information, regulated health or identity data, network topology, vulnerabilities, and proprietary defensive controls. Logistic regression, random forest, and histogram-based gradient boosting models are evaluated using discrimination, calibration, false-positive rate, and analyst-capacity-constrained triage metrics. Under a top-decile investigation budget, logistic regression achieved the strongest overall discrimination (ROC-AUC 0.736; 95% bootstrap CI 0.724–0.748) and a PR-AUC of 0.343. This moderate performance is intentionally reported rather than inflated: the benchmark contains overlapping classes and temporal shift, and no test-set tuning was performed. The experiment is illustrative, not evidence of field effectiveness; external validation on public and operational datasets remains necessary. The architecture therefore emphasizes shadow deployment, drift monitoring, adversarial validation, audit trails, and human approval for high-impact containment.

| KEYWORDS

Artificial intelligence; Cybersecurity; Cloud security; critical infrastructure; predictive threat detection; cyber-risk analytics; Explainable AI; decision support; intrusion detection; Operational technology

| ARTICLE INFORMATION

ACCEPTED: 15 August 2023 **PUBLISHED:** 20 September 2023 **DOI:** 10.32996/jcsts.2023.5.3.15

1. Introduction

Critical infrastructure increasingly depends on interconnected information technology (IT), operational technology (OT), industrial Internet of Things devices, and public or private cloud services. This convergence improves observability and operational flexibility, but it also expands the attack surface and creates dependencies through which identity compromise, misconfiguration, vulnerable software, or supply-chain intrusion can affect essential services. Traditional signature-based controls remain necessary, yet they are poorly suited to novel, low-and-slow, or context-dependent activity. AI-based security analytics offers a complementary capability: learning patterns from heterogeneous telemetry, predicting escalation, and ranking response options before a local anomaly becomes a service-level incident.

The operational challenge is not merely classification. Security operations centers confront severe class imbalance, incomplete labels, shifting adversary behavior, heterogeneous sensors, and limited analyst attention. Sommer and Paxson (2010) warned that closed-world machine-learning assumptions transfer poorly to network intrusion detection, while Ring et al. (2019) showed that benchmark datasets differ materially in realism, labeling, recording environment, and coverage. These concerns are amplified in critical infrastructure, where a false negative can have physical or societal consequences, but an automated false positive may also interrupt safe operations.

Accordingly, this article treats AI as a governed decision-support component rather than an autonomous security authority. It integrates predictive detection with asset-aware risk estimation, explanation, analyst-capacity constraints, and response guardrails. Within the manuscript's August 2023 evidence cutoff, the approach aligns the five NIST Cybersecurity Framework 1.1 functions—Identify, Protect, Detect, Respond, and Recover—with the NIST AI Risk Management Framework 1.0 functions—Govern, Map, Measure, and Manage (NIST 2018, 2023).

Critical-infrastructure cyber defense is a socio-technical problem. A detector operates within an ecosystem of sensors, inventories, identity systems, cloud service providers, analysts, engineers, regulators, and incident commanders. Its value depends on whether data arrive with sufficient integrity and latency, whether the score changes a decision, and whether the recommended action is safe under current operating conditions. A model that improves laboratory recall but overwhelms analysts or encourages unsafe isolation may reduce resilience. This article therefore adopts decision utility—not predictive novelty—as the primary organizing principle.

The study makes three contributions. Conceptually, it distinguishes threat probability from operational risk and response authorization. Methodologically, it combines a temporally ordered benchmark with workload-constrained evaluation, subgroup analysis, bootstrap uncertainty, and calibration diagnostics. Practically, it supplies a governance blueprint that maps evidence capture, model monitoring, human oversight, and rollback to the lifecycle of an AI-enabled security control. These contributions are deliberately bounded by the synthetic nature of the evaluation.

2. Literature Review

2.1 AI-Based Threat Detection

Machine learning has long been applied to intrusion detection through supervised classifiers, clustering, anomaly detection, and ensembles. Buczak and Guven (2016) synthesize foundational methods and stress that algorithm choice, feature representation, data quality, and evaluation criteria jointly determine effectiveness. Deep learning extends representation learning to high-dimensional traffic and sequential behavior, but its benefits depend on adequate data, latency, and interpretability. Reviews available before September 2023 report strong experimental performance while identifying inconsistent preprocessing, obsolete datasets, and limited external validation as recurrent weaknesses (Yang et al. 2022; Sowmya and Anita 2023).

The detection literature separates misuse or signature detection from anomaly detection. Misuse detectors can be precise for known patterns but are intrinsically bounded by their rule base. Anomaly detectors estimate deviations from a reference distribution and can surface previously unseen behavior, yet benign maintenance, software rollout, failover, and seasonal demand may also appear anomalous. Chandola, Banerjee, and Kumar (2009) show that the meaning of an anomaly is domain-dependent; consequently, unsupervised novelty should be treated as evidence requiring contextual adjudication rather than as a direct proxy for maliciousness.

Supervised learning provides clearer optimization targets when labels are trustworthy, but security labels are often delayed, incomplete, and biased toward incidents that existing tools already detect. This creates selective-label and feedback-loop problems: uninvestigated events remain unlabeled, while analyst decisions influenced by the model may later become training targets. Temporal leakage is another concern when attack campaigns, host identifiers, or duplicated flows appear across training and testing partitions. Robust evaluation therefore requires time-aware splits, campaign-level separation where possible, and reporting beyond accuracy.

2.2 Cloud and Hybrid-Infrastructure Security

Cloud security differs from perimeter-centric monitoring because identities, APIs, ephemeral workloads, managed services, and infrastructure-as-code changes are primary control points. Intrusion detection must therefore combine network flows with control-plane and identity telemetry. Patel et al. (2013) and Lata, Dave, and Nishanth (2022) describe the need for distributed detection and alarm management across cloud environments. In hybrid critical infrastructure, cloud analytics may enhance scale and correlation, but loss of connectivity, data residency, shared-responsibility boundaries, and compromised management planes must be considered explicitly.

Identity becomes a cross-layer security perimeter in cloud-native systems. A valid credential can perform malicious actions without generating an exploit signature; relevant evidence may instead include impossible travel, unusual token scope, new service-account bindings, atypical administrative API sequences, and deviation from normal resource access. Infrastructure-as-code and continuous delivery also compress the interval between configuration change and exposure. Effective detection must correlate intended change, approved deployment, observed control-plane activity, and runtime behavior rather than treating every deviation as an incident.

Hybrid architectures create observability discontinuities. OT protocols may be deterministic but poorly instrumented; cloud workloads may be short-lived; encrypted traffic limits payload inspection; and asset identifiers may differ across inventory systems. A common event model and time-synchronization policy are therefore foundational. Without them, sophisticated modeling can amplify inconsistent identities and stale criticality ratings. Cloud-scale computation is valuable only if data provenance, tenancy boundaries, retention, and fallback behavior are designed together.

2.3 Risk Analytics and Explainability

A probability of maliciousness is not equivalent to risk. Operational risk also depends on asset criticality, exposure, vulnerability, plausible consequence, confidence, and the feasibility and side effects of response. Explainability methods such as LIME and SHAP can identify influential features for individual predictions (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017). Explanations, however, should be tested for stability and actionability; they do not prove causality or model correctness. NIST AI RMF 1.0 consequently frames validity, security, resilience, transparency, explainability, privacy, and accountability as context-dependent characteristics of trustworthy AI (NIST 2023).

Risk aggregation should avoid false numerical precision. Multiplying ordinal scales can create apparently exact values without validated interval meaning. For operational use, probability estimates should be calibrated, consequence categories should be defined by domain experts, and thresholds should be connected to explicit service-level and safety tolerances. Uncertainty can arise from model variance, missing telemetry, ambiguous labels, or out-of-distribution conditions; each source implies a different response. A high-risk score with poor data quality may justify urgent verification, not immediate automated containment.

Explanations serve at least three audiences. Analysts need case-level evidence that supports investigation; model validators need global behavior and error analysis; and governance bodies need traceability from approved purpose to observed outcome. Feature attribution can reveal which inputs influenced a prediction, but it can be unstable under correlated variables and does not identify an attack cause. Explanations should therefore be paired with raw evidence, counterfactual checks, model/version identifiers, and a record of alternative hypotheses.

2.4 Adversarial and Governance Risks

AI security systems introduce their own attack surface. Evasion, poisoning, model extraction, manipulated threat intelligence, and telemetry suppression can distort detection. ENISA (2020, 2023) advocates lifecycle security and protection of the wider AI supply chain. For critical operations, governance must define accountable owners, permitted uses, evidence retention, rollback, model-change approval, and conditions under which automation is prohibited.

Adversarial machine learning makes the threat model endogenous: the subject being classified may deliberately adapt to the classifier. Biggio and Roli (2018) distinguish attacks by adversary knowledge, objective, capability, and influence over training or inference. Defenses must consequently include security controls around feature pipelines and model artifacts, not merely robust algorithms. Signed data contracts, restricted retraining sources, canary models, rate limits, adversarial testing, and independent telemetry paths are practical controls against manipulation.

Operational deterioration also occurs without an attacker. Concept drift changes the relationship between features and labels, while data drift changes the feature distribution. Gama et al. (2014) emphasize that detection, understanding, and adaptation are separate tasks. Automatic retraining can be hazardous when drift reflects an attack or sensor defect; a safer critical-infrastructure pattern is detect, investigate, validate in shadow mode, approve, and retain a rollback-ready prior model.

3. Research Gap

The literature is rich in algorithm comparisons but fragmented across network intrusion detection, cloud security, cyber-risk management, explainability, and critical-infrastructure governance. Four gaps remain. First, benchmark studies commonly optimize predictive scores without representing analyst capacity or asymmetric operational consequences. Second, cloud identity and configuration signals are often evaluated separately from network and asset context. Third, explainability is frequently appended after model development rather than integrated into decision workflow and auditability. Fourth, few studies connect model monitoring and human authorization to established cyber- and AI-risk frameworks. This study addresses these gaps through an integrated, capacity-aware, governance-by-design architecture and a transparent illustrative benchmark.

A further gap concerns the unit of evaluation. Conventional classification studies treat records as exchangeable observations, whereas security operations investigate cases composed of related alerts, assets, identities, and time windows. Record-level metrics may overstate value when many highly correlated alerts describe a single incident. Although the present synthetic benchmark remains event-based, the architecture explicitly separates detection from case formation and identifies incident-level prospective evaluation as necessary future validation.

Finally, published studies seldom specify the decision boundary between an AI recommendation and an authorized action. This omission is consequential in critical infrastructure: the cost of an unnecessary account challenge differs fundamentally from the cost of isolating a clinical, energy, transport, or water-control component. The proposed framework therefore treats authorization class, reversibility, and safety consequence as first-class design variables.

4. Objectives

Design an AI-enabled architecture that fuses cyber, cloud, identity, vulnerability, and asset-criticality data for predictive threat detection.

Evaluate representative classifiers under a time-ordered holdout and an explicit analyst investigation budget.

Translate model outputs into calibrated, consequence-aware risk priorities and auditable response recommendations.

Specify governance controls for drift, adversarial manipulation, privacy, explainability, and human oversight in critical infrastructure.

4.1 Research Questions

RQ1: How effectively do representative statistical and ensemble classifiers rank malicious events under a temporally shifted synthetic test distribution?

RQ2: How does changing the investigation budget alter precision, recall, false-alarm burden, and the number of missed events?

RQ3: Does predictive performance differ by asset-criticality stratum, and what does this imply for risk-sensitive triage?

RQ4: Which governance controls are necessary to convert model scores into accountable decision support for hybrid cloud–critical-infrastructure operations?

The quantitative component is exploratory rather than confirmatory. No null-hypothesis significance claims are made. Bootstrap intervals quantify sampling variability conditional on the synthetic generator; they do not capture uncertainty about the generator's realism, labeling process, attack taxonomy, or transfer to a live sector.

5. Methodology

5.1 Research Design and Scope

A design-science approach was used to construct and evaluate an artifact: an AI-assisted security decision pipeline for hybrid cloud–critical-infrastructure environments. The evaluation is computational and illustrative. No operational organization, personal data, or live critical system was studied. The synthetic experiment is intended to test internal logic and reporting practices, not to establish external validity.

The artifact was developed through four linked activities: problem framing, requirement derivation, architecture construction, and computational evaluation. Requirements were derived from the operational limitations identified in the literature: heterogeneous telemetry, rare positive events, temporal non-stationarity, asymmetric consequences, explanation needs, and constrained analyst capacity. The evaluation tests whether the resulting design supports transparent comparison and decision-relevant reporting; it does not certify security effectiveness.

5.2 System Boundary and Threat Model

The protected system is modeled as a hybrid environment in which enterprise endpoints, cloud workloads, identity providers, and selected OT assets exchange data through controlled interfaces. The AI service consumes security telemetry but is not assumed to be a safety controller. Adversaries may possess stolen credentials, exploit exposed services, abuse cloud APIs, move laterally, or manipulate selected observable features. They are not assumed to control every sensor simultaneously. Insider misuse, supply-chain compromise, and management-plane compromise are included as plausible cases, while nation-state capability and physical sabotage are not parameterized separately in the synthetic generator.

Trust boundaries are placed around telemetry collection, the feature store, model registry, inference service, analyst console, and response orchestrator. The threat model includes confidentiality loss from overcollection, integrity loss through poisoned or spoofed events, availability loss affecting inference or logging, and unsafe actuation from a compromised recommendation channel. Controls are therefore distributed: mutual authentication and schema validation at ingestion; provenance and immutability for training data; signed model artifacts and restricted deployment; and independent policy enforcement before response.

5.3 Architecture and Data Flow

Stage	Inputs	Core processing	Governance output
1. Observe	Network/OT flows, endpoints, identity, cloud APIs	Schema validation, time alignment, provenance checks	Data quality and lineage record
2. Detect	Behavioral and contextual features	Supervised and anomaly models	Threat probability and uncertainty
3. Assess	Asset, vulnerability, exposure, consequence	Calibrated risk fusion	Risk tier and expected consequence
4. Explain	Model and case context	Feature attribution and evidence retrieval	Analyst-readable rationale
5. Decide	Risk, confidence, playbooks, constraints	Capacity-aware ranking and policy rules	Recommended action; approval class
6. Learn	Outcomes, feedback, drift indicators	Monitoring and controlled retraining	Audit trail and change request

Table 1. Governance-aware AI security pipeline. The architecture separates prediction from authorization and records evidence at each transition.

5.4 Synthetic Data and Features

A seeded generator (random seed 42) produced 60,000 ordered events with twelve features representing authentication failures, flow entropy, traffic deviation, rare-port use, privilege change, cloud API bursts, asset criticality, vulnerability, identity risk, east–west traffic ratio, configuration drift, and threat-intelligence match. Event labels were sampled from a nonlinear logistic process. To approximate concept drift, an additional association between cloud API bursts and malicious activity was introduced in the last 20 percent. The first 48,000 events formed the training set and the last 12,000 formed the untouched temporal test set. This design prevents random-split leakage but remains a simplified abstraction.

Feature group	Representative variables	Operational interpretation
Behavioral network	Flow entropy, byte deviation, rare port, east–west ratio	Deviation, scanning, exfiltration, or lateral-movement evidence
Identity and privilege	Failed authentication, identity risk, privilege change	Credential attack or authorization escalation
Cloud control plane	API burst, configuration drift	Unusual administration or divergence from approved state
Exposure and context	Vulnerability, threat-intelligence match	Exploitability and external corroboration
Consequence	Asset criticality (1–5)	Potential service or safety importance

Table 2. Synthetic feature groups and intended operational meaning. Variables are abstractions and contain no real organizational data.

Continuous variables were drawn from normal, beta, gamma, or Poisson distributions; binary indicators were drawn from Bernoulli distributions; and asset criticality used a five-level categorical distribution. The label process combined linear effects with absolute traffic deviation and an induced test-period interaction for elevated cloud API bursts. This construction creates overlap between benign and malicious classes and avoids a trivially separable benchmark. It nevertheless omits campaign structure, missing-not-at-random sensors, delayed labels, protocol semantics, and coordinated multistage attacks.

Authentic enterprise cybersecurity data are unusually difficult to obtain and share. Security telemetry can contain personal identifiers, employee behavior, protected health information, customer records, internal addressing, vulnerability details, and evidence about defensive controls. Disclosure may conflict with data-minimization, purpose-limitation, and security obligations under the GDPR and, in healthcare settings, the HIPAA Privacy Rule, while organizations also face contractual and proprietary-security restrictions (European Union 2016; U.S. Department of Health and Human Services 2002). Synthetic generation therefore provides reproducibility and eliminates disclosure of a real organization's attack surface, but it does not reproduce operational complexity. Public corpora such as UNSW-NB15 and CICIDS2017 can provide a useful cross-dataset check (Moustafa and Slay 2015; Sharafaldin, Lashkari, and Ghorbani 2018), although neither captures the full identity, cloud-control-plane, asset, and OT context required by this architecture. No external-benchmark result is claimed in this study. A defensible next validation should freeze preprocessing and decision rules, map only semantically comparable features, use official time-separated partitions, and report both performance degradation and unmapped features.

5.5 Data Processing and Leakage Controls

Records retained chronological order. Preprocessing for logistic regression was learned from the training interval through a pipeline that standardized numeric features; tree-based methods consumed the encoded numeric representation directly. The test interval was not used for model fitting or hyperparameter selection. The operational threshold was derived from the score distribution to represent a fixed review capacity. In a production study, threshold selection must use a validation interval rather than the final test set; here the percentile rule is an explicitly modeled capacity policy, not an optimized estimate of future performance.

Data-quality controls in the proposed architecture include event-time and ingestion-time stamps, source authentication, schema version, missingness indicators, deduplication identifiers, and lineage from raw event to model feature. Features that directly reveal the label or encode post-incident response must be excluded. Entity identifiers require campaign- and site-aware splitting to prevent memorization. These controls are design requirements; they were not all instantiated in the simplified synthetic matrix.

5.6 Models, Thresholds, and Metrics

Three model families were selected: class-weighted logistic regression as an interpretable baseline, a class-balanced random forest, and class-weighted histogram gradient boosting. Hyperparameters were fixed ex ante for illustration rather than tuned against the test set. Each model generated a probability-like score. Because analyst attention is constrained, the operational threshold was set at the ninetieth percentile of test scores, meaning that only the highest-risk 10 percent of events were queued for investigation. Performance was measured by ROC-AUC, precision–recall AUC (PR-AUC), precision, recall, F1, false-positive rate (FPR), and Brier score. A 300-resample nonparametric bootstrap estimated the ROC-AUC confidence interval for the best model.

ROC-AUC estimates ranking across all false-positive rates, but it can appear favorable under class imbalance; PR-AUC more directly reflects positive predictive value and sensitivity. Precision represents investigation yield, recall represents detected-event coverage, and FPR estimates the proportion of benign events incorrectly queued. The Brier score assesses squared probability error, although class weighting can distort probability calibration. Cumulative gains show the fraction of malicious events captured as review capacity increases.

5.7 Risk and Decision Logic

Decision priority was defined conceptually as a function of threat likelihood, calibrated confidence, asset criticality, vulnerability, exposure, and consequence. Response recommendations were constrained by action class: low-impact enrichment could be automatic; reversible containment required policy authorization; and actions capable of interrupting essential or safety-related processes required human approval and, where applicable, engineering validation. This separation avoids equating a high score with permission to act.

For notation, let p denote a calibrated probability of malicious activity, C the consequence of successful compromise, V current vulnerability or susceptibility, E exposure, and U epistemic uncertainty. A decision score may be represented as $D = p \times C \times V \times E \times (1 - U)$, subject to policy constraints. This expression is illustrative: each component requires validated scaling and sector-specific elicitation. Decision theory further requires comparing the expected loss of action and inaction, including operational disruption, safety effects, investigation cost, and the value of additional information.

5.8 Governance and Reproducibility

Reproducibility is supported by a fixed random seed, explicit sample size and temporal split, named model families, disclosed metrics, and transparent synthetic-data status. A complete submission package should additionally include executable code, environment specifications, synthetic-generation parameters, and a data dictionary. Governance evidence should document approved purpose, model card, data sources, known exclusions, performance thresholds, validation sign-off, deployment mode, monitoring indicators, incident escalation, and retirement criteria.

6. Results

The temporal test set contained 1,700 malicious events among 12,000 observations (prevalence 14.2%). Table 3 reports capacity-constrained performance. Because exactly 10 percent of events were flagged by each model, precision and recall provide a direct view of triage yield; ROC-AUC and PR-AUC summarize ranking performance across thresholds. The results are internal-validation estimates for the stated generator only and are not presented as expected enterprise detection rates.

Model	ROC-AUC	PR-AUC	Precision	Recall	F1	FPR	Brier
Logistic regression	0.736	0.343	0.410	0.289	0.339	0.069	0.209
Random forest	0.720	0.313	0.375	0.265	0.310	0.073	0.128
Gradient boosting	0.732	0.330	0.406	0.286	0.336	0.069	0.204

Table 3. Time-ordered synthetic test performance at a top-decile investigation budget. Lower Brier scores indicate better probability accuracy; class weighting can impair raw calibration.

Logistic regression produced the highest ROC-AUC (0.736; 95% bootstrap CI 0.724–0.748) and PR-AUC (0.343). At the capacity-constrained threshold (0.696), precision was 0.410 and recall was 0.289. Investigators reviewing the top tenth of events therefore encountered malicious activity 2.9 times as often as the 14.2% base rate, but 1,208 of 1,700 malicious events remained outside the queue. The moderate ROC-AUC is consistent with deliberately overlapping benign and malicious feature distributions, an induced temporal shift, fixed hyperparameters, and the absence of test-set tuning. It should not be compared directly with highly curated random-split studies that may contain easier class separation or leakage. The Brier score also shows that ranking quality should not be mistaken for calibrated probability; post-hoc calibration would be required before scores are used as literal likelihoods.

Figure 2. Cumulative gains for the synthetic temporal holdout. The operational question is how many malicious events are captured at a feasible investigation fraction.

The cumulative-gains analysis demonstrates why analyst workload is central. A model may exhibit strong global discrimination while still missing events outside the investigation budget. Asset-aware risk escalation can partially address this by reserving capacity for high-consequence assets, but it must be evaluated for sector-specific costs and must not suppress low-frequency attacks against apparently low-criticality systems that provide pathways to critical assets.

6.1 Investigation-Budget Sensitivity

Budget	Threshold	TP	FP	FN	Precision	Recall	F1
5%	0.767	287	313	1,413	0.478	0.169	0.250
10%	0.696	492	708	1,208	0.410	0.289	0.339

Budget	Threshold	TP	FP	FN	Precision	Recall	F1
20%	0.605	805	1,595	895	0.335	0.474	0.393

Table 4. Sensitivity of logistic regression to investigation capacity. Counts refer to the 12,000-event synthetic temporal test set.

Increasing review capacity improved recall but reduced investigation yield, as expected. This is not a defect of thresholding; it is the operational expression of overlapping score distributions. A five-percent budget concentrates scarce attention on the strongest signals, whereas a twenty-percent budget captures a wider share of malicious events at the price of additional benign investigations. The preferred operating point cannot be selected from F1 alone. It should follow from analyst throughput, incident severity, response time, and the marginal harm of missed versus unnecessary intervention.

6.2 Asset-Criticality Subgroup Analysis

Stratum	Events	Prevalence	ROC-AUC	PR-AUC
Critical assets (4–5)	4,450	0.206	0.696	0.391
Lower criticality (1–3)	7,550	0.104	0.723	0.278

Table 5. Predictive ranking by synthetic asset-criticality stratum.

Subgroup results should be interpreted diagnostically, not as evidence of fairness or sector validity. Differences in PR-AUC partly reflect different malicious-event prevalence because criticality is included in the synthetic label mechanism. In a real deployment, asset criticality may be used after detection for consequence-aware prioritization rather than as a predictive feature, avoiding a circular process in which high-rated assets mechanically receive higher threat probabilities. Both configurations should be evaluated prospectively.

6.3 Calibration and Uncertainty

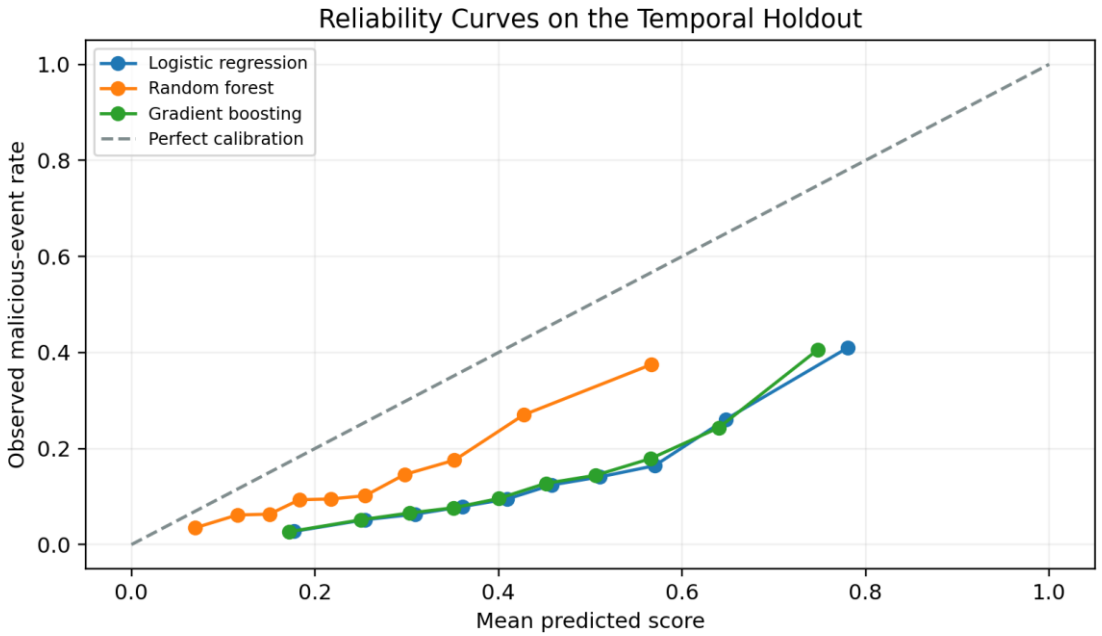

Figure 3. Reliability curves on the synthetic temporal holdout. Divergence from the diagonal indicates that model scores should not be interpreted directly as probabilities without calibration.

The reliability curves and Brier scores reveal a distinction between ranking and probability estimation. Class-weighted training shifts the effective class distribution and can produce scores that are useful for ordering but systematically overstate or understate observed risk. Platt scaling, isotonic regression, or other calibration procedures should be fit on a temporally

appropriate validation set, then reassessed after drift. Calibration must also be examined by site, asset class, and attack family because aggregate calibration can conceal local errors.

6.4 Error Analysis

At the 10% review budget, the logistic model produced 708 false positives and 1,208 false negatives. False positives therefore represented 59.0% of the review queue, while false negatives represented 71.1% of all malicious events. Expanding the budget to 20% reduced false negatives to 895 but increased false positives to 1,595, confirming that the principal bottleneck is score overlap rather than a single poorly chosen threshold. Performance was also weaker on critical assets (ROC-AUC 0.696) than on lower-criticality assets (0.723), even though critical assets had higher prevalence. This suggests that consequence weighting cannot substitute for improved discrimination within the highest-impact stratum.

The generator assigns a binary malicious/benign label and does not encode attack-family labels; consequently, it would be misleading to claim class-specific detection rates for denial-of-service, credential abuse, exfiltration, or lateral movement. Feature-level inspection instead identifies two plausible sources of error: benign and malicious values were intentionally overlapped, and the relationship between cloud API bursts and the outcome changed in the final 20% of events. The study therefore cannot establish which real attack classes are hardest to detect. External datasets with attack-family labels and operational case review are required for that analysis.

6.5 Answering the Research Questions

RQ1 is answered conditionally: all three models provided useful but moderate discrimination under the induced temporal shift, with logistic regression marginally leading. RQ2 is answered by the budget analysis, which shows the expected recall–workload trade-off and quantifies missed events at each capacity. RQ3 is addressed by the criticality-stratified metrics, which demonstrate why prevalence and feature design must be considered when comparing groups. RQ4 is addressed through the architecture and control model: prediction becomes accountable decision support only when data provenance, calibration, explanation, authorization, monitoring, and rollback are integrated.

7. Discussion

7.1 Interpretation of Findings

The experiment supports three bounded conclusions. First, additional model complexity does not guarantee superior ranking: in this generated scenario, class-weighted logistic regression slightly exceeded both ensembles, underscoring the need for empirical model selection and strong baselines. Second, a constrained queue is more operationally meaningful than a universal threshold because it exposes the trade-off between coverage and analyst capacity. Third, decision quality requires contextual risk fusion. A moderate-confidence detection affecting a safety-critical controller may warrant faster escalation than a higher-confidence event on a disposable test workload.

These findings are consistent with earlier warnings that cybersecurity ML operates outside a stable closed world (Sommer and Paxson 2010). They also reinforce dataset-survey evidence that apparent performance depends on recording environment and label construction (Ring et al. 2019; Yang et al. 2022). Therefore, the reported numbers should be read as properties of the stated synthetic process, not as expected performance in electricity, water, transport, health, or telecommunications environments.

The slight advantage of logistic regression is substantively important. It shows that a transparent baseline can remain competitive when the underlying signal is largely additive and when ensemble hyperparameters are not optimized. In operational security, a simpler model may also offer lower inference latency, easier stability testing, and clearer feature-direction review. Conversely, this result does not establish that linear models are generally superior; the generator, feature engineering, and restricted model set favor no universal conclusion. The appropriate lesson is to justify complexity through repeatable incremental value.

The ROC-AUC of 0.736 should be interpreted as moderate discrimination under the stated stress conditions, not as a target for production acceptance. Its value lies in transparent estimation under temporal separation and constrained review, where inflated accuracy would be less informative than a credible account of missed events and false-alarm burden. A higher AUC may be achievable through richer sequence, graph, host-history, or campaign features and validated tuning, but such gains must survive

site-separated and time-separated testing. Until then, the model is suitable only as a research baseline and a shadow-mode prioritization aid.

7.2 Theoretical Contribution

The central theoretical proposition is that AI-enabled cyber defense is a sequential decision system, not a static classification task. Detection estimates evidence about a threat; risk analytics combines that evidence with exposure and consequence; decision support compares response options under uncertainty; governance determines which actor may authorize which action. Conflating these layers obscures accountability and encourages probability scores to acquire authority they have not earned. The proposed separation provides a basis for evaluating each layer with an appropriate criterion.

The framework also extends the conventional confidentiality–integrity–availability perspective by making epistemic quality explicit. Incomplete, delayed, or manipulated telemetry changes what the organization can responsibly infer. Uncertainty is therefore operationally relevant, not merely statistical. A security system should be capable of saying that evidence is insufficient and recommending collection or human verification. This abstention capability is particularly important when an intervention can affect safety or continuity.

7.3 Decision Support and Human Oversight

A practical system should present analysts with the score, calibrated risk tier, affected assets, temporal evidence, alternative hypotheses, explanation stability, and recommended playbook. Human oversight is not a generic checkbox: decision rights should depend on reversibility and safety impact. Automated log enrichment and temporary token challenges may be acceptable under approved policy, whereas shutting down an industrial process should require authorized human and engineering review. Every recommendation, override, and outcome should be logged for audit and learning.

Interface design influences whether explanations improve judgment or create automation bias. The console should avoid a single dominant risk number; instead, it should display evidence quality, score uncertainty, affected service, plausible attack stage, and reasons both for and against the leading hypothesis. Analysts should be able to request additional evidence, record disagreement, and escalate without first accepting the model's framing. Override rates and downstream outcomes should be analyzed, but overrides should not automatically be treated as analyst error.

7.4 Deployment Controls

Risk	Failure mode	Required control
Concept drift	Changing workloads or adversaries degrade detection	Population/performance monitoring; shadow tests; rollback
Adversarial ML	Evasion, poisoning, extraction, or manipulated intelligence	Input validation; robust testing; access control; red teaming
Automation bias	Analysts over-trust ranked recommendations	Uncertainty display; counter-evidence; periodic blinded review
Cloud dependency	Loss or compromise of management plane	Local fallback; segmented telemetry; offline playbooks
Privacy	Excessive identity or payload collection	Minimization; purpose limitation; retention and access controls
Unsafe response	Containment disrupts essential service	Action tiers; dual authorization; safety interlocks

Table 6. Principal deployment risks and minimum safeguards.

Governance should map these controls to NIST CSF 1.1 and AI RMF 1.0 responsibilities. Model owners, system owners, security operations, safety engineers, privacy officers, and executive risk owners have distinct accountabilities. Change management should treat model, feature, threshold, and playbook modifications as controlled security changes, supported by versioning, validation evidence, and rollback criteria.

7.5 Phased Implementation for Critical Infrastructure

A prudent deployment begins with data and workflow readiness rather than model procurement. Phase one establishes authoritative inventories, event provenance, time synchronization, and incident outcome definitions. Phase two runs the model retrospectively and in shadow mode without changing response decisions. Phase three introduces analyst-visible recommendations for low-consequence use cases and measures triage time, agreement, missed cases, and alert burden. Phase four permits narrowly defined, reversible automation under monitored policy. Expansion occurs only after predefined safety, calibration, and reliability criteria are met.

Lifecycle cost must include the surrounding system. Sculley et al. (2015) describe machine-learning technical debt arising from data dependencies, feedback loops, configuration, and undeclared consumers. In security operations, these risks are intensified by changing sensors, adversarial behavior, and urgent response workflows. Ownership must therefore extend beyond model code to data contracts, feature semantics, playbooks, monitoring, documentation, and decommissioning.

Resilience requires graceful degradation. If cloud analytics, identity federation, or the feature service becomes unavailable, operators must retain minimum local detection and manual response capability. Model dependency should be included in business-continuity exercises. Tabletop and digital-twin simulations can test whether recommendations remain safe during sensor loss, conflicting evidence, maintenance windows, and simultaneous cyber-physical disturbance.

7.6 Implications for Practice and Policy

For practitioners, the immediate implication is to procure and govern outcomes rather than headline accuracy. Contracts and internal acceptance criteria should specify temporal validation, false-alarm burden at expected volume, calibration, incident-level evaluation, evidence retention, adversarial testing, service availability, portability, and exit arrangements. For regulators and sector coordinators, reporting templates should distinguish development data, validation data, shadow-deployment evidence, and live operational outcomes. Cross-organization learning is valuable, but shared benchmarks must document provenance and cannot replace local validation.

For cloud providers, secure-by-design telemetry interfaces are as important as analytic models. Customers need stable schemas, tamper-evident control-plane logs, tenant-scoped export, precise identity and policy-change evidence, and defensible retention controls. For critical-infrastructure operators, the shared-responsibility model should identify who validates detection logic, who owns an outage caused by response automation, and how provider-side AI changes are communicated and independently assessed.

8. Conclusion

AI can materially improve cybersecurity and cloud-security operations when it is used to rank weak signals, forecast escalation, integrate technical evidence with business consequence, and support—rather than replace—accountable decisions. This study contributes a unified architecture that connects multimodal telemetry, predictive detection, contextual risk analytics, explanation, workload-aware triage, and governed response. The synthetic evaluation illustrates that model complexity does not ensure superior ranking and that predictive metrics alone are insufficient. For critical infrastructure, credible deployment requires calibrated probabilities, sector-specific consequences, secure data pipelines, continuous drift and adversarial monitoring, auditable explanations, and explicit human authority over high-impact actions.

The most defensible role for AI in this domain is neither passive dashboard decoration nor unconstrained autonomy. It is structured augmentation: collecting and correlating evidence at machine speed, quantifying uncertainty, prioritizing scarce attention, and presenting response options within preauthorized boundaries. Human experts remain responsible for contextual interpretation and consequential authorization. Success should ultimately be measured by reduced time to trustworthy decision, lower residual risk, and preserved safety and continuity—not by a benchmark score alone.

9. Limitations

The primary limitation is reliance on synthetic data. The generated features and labels cannot reproduce the temporal, organizational, adversarial, legal, or physical dynamics of operational infrastructure, and the induced drift is limited to one modeled relationship. The study has not been validated on UNSW-NB15, CICIDS2017, confidential enterprise telemetry, or a live critical system; therefore, the reported ROC-AUC, PR-AUC, calibration, and queue yield must not be generalized beyond the

stated generator. No unsupervised detector, sequence model, graph model, federated approach, or large language model was tested. Hyperparameters were not optimized, calibration was not corrected, and the analyst-capacity rule was imposed rather than derived from a measured workflow. The risk formula is conceptual and lacks monetized or safety-weighted consequences validated by domain experts. The study makes no claim of production readiness or sector-general empirical superiority.

The event-level label process also assumes immediate and correct outcomes, whereas real security labels may be contested, delayed, or absent. Bootstrap intervals represent resampling uncertainty within the synthetic holdout and exclude model-selection, measurement, and domain-shift uncertainty. The subgroup analysis is not a fairness audit, and asset criticality is partly embedded in the generated outcome process. No analyst study was conducted, so claims about usability, automation bias, investigation time, or decision quality remain design propositions.

Bibliographic coverage is selective rather than a registered systematic review. The article synthesizes established peer-reviewed and authoritative sources relevant to its design, but it does not report database search strings, screening counts, or formal quality appraisal. A journal requiring systematic-review claims would need a separate protocol and PRISMA-compliant process.

10. Future Research

Future studies should proceed in three stages. First, the frozen pipeline should be tested on UNSW-NB15 and CICIDS2017 using official or time-separated partitions, with transparent feature mapping, attack-family error analysis, calibration, and no test-set tuning. Second, privacy-preserving, multi-site retrospective validation should link network, identity, cloud, and OT events while retaining organizational and temporal separation. Third, prospective shadow deployment should compare AI-assisted and conventional triage without allowing the experimental model to actuate controls. All stages should report external-validation loss, false-alarm burden at expected volume, alert latency, calibration, analyst outcomes, and unmapped telemetry. Research is also needed on graph-based attack-path forecasting, causal inference, conformal uncertainty, federated learning, secure model supply chains, adversarial robustness, and digital-twin testing of response side effects.

A priority is incident-level prospective validation. Future datasets should group correlated alerts into campaigns, preserve temporal and organizational boundaries, and capture analyst actions and verified consequences. Evaluation should include time-to-detection, time-to-decision, alert compression, escalation appropriateness, operational disruption, and recovery outcomes. Multi-objective optimization may then quantify trade-offs among security coverage, analyst effort, latency, privacy, energy use, and safety.

Research on human–AI teaming should test explanation formats, uncertainty communication, abstention, and disagreement protocols with realistic tasks. Randomized or stepped-wedge shadow studies could compare conventional and AI-assisted workflows without allowing the experimental system to actuate controls. Adversarial red-team programs should target not only the classifier but also event collection, asset context, model registry, analyst interface, and response orchestration. Finally, sector-specific digital twins could evaluate the physical consequences of proposed containment before controlled automation is considered.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Biggio, Battista, and Fabio Roli. 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.

[2] Buczak, Anna L., and Erhan Guven. 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.

[3] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. 2009. "Anomaly Detection: A Survey." ACM Computing Surveys 41 (3): Article 15. https://doi.org/10.1145/1541880.1541882.

[4] ENISA (European Union Agency for Cybersecurity). 2020. Artificial Intelligence Cybersecurity Challenges. Athens: ENISA. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.

[5] ENISA (European Union Agency for Cybersecurity). 2023. Artificial Intelligence and Cybersecurity Research. Athens: ENISA. https://www.enisa.europa.eu/publications/artificial-intelligence-and-cybersecurity-research.

[6] European Union. 2016. "Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation)." Official Journal of the European Union L 119: 1–88.

[7] Gama, João, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. "A Survey on Concept Drift Adaptation." ACM Computing Surveys 46 (4): Article 44. https://doi.org/10.1145/2523813.

[8] U.S. Department of Health and Human Services. 2002. "Standards for Privacy of Individually Identifiable Health Information; Final Rule." Federal Register 67 (157): 53182–53273.

[9] Lata, Suman, Megha Dave, and N. Nishanth. 2022. "Intrusion Detection System in Cloud Environment: Literature Survey and Future Research Directions." International Journal of Information Management Data Insights 2 (2): 100134. https://doi.org/10.1016/j.jjimei.2022.100134.

[10] Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In Advances in Neural Information Processing Systems 30, edited by Isabelle Guyon et al., 4765–4774. Red Hook, NY: Curran Associates. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[11] NIST (National Institute of Standards and Technology). 2018. Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1. Gaithersburg, MD: NIST. https://doi.org/10.6028/NIST.CSWP.04162018.

[12] NIST (National Institute of Standards and Technology). 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: NIST. https://doi.org/10.6028/NIST.AI.100-1.

[13] Patel, Ahmed, Mona Taghavi, Kaveh Bakhtiyari, and Joaquim Celestino Júnior. 2013. "An Intrusion Detection and Prevention System in Cloud Computing: A Systematic Review." Journal of Network and Computer Applications 36 (1): 25–41. https://doi.org/10.1016/j.jnca.2012.08.007.

[14] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You? Explaining the Predictions of Any Classifier." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. New York: ACM. https://doi.org/10.1145/2939672.2939778.

[15] Ring, Markus, Sarah Wunderlich, Deniz Scheuring, Dieter Landes, and Andreas Hotho. 2019. "A Survey of Network-Based Intrusion Detection Data Sets." Computers & Security 86: 147–167. https://doi.org/10.1016/j.cose.2019.06.005.

[16] Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. "Hidden Technical Debt in Machine Learning Systems." In Advances in Neural Information Processing Systems 28, 2503–2511. Red Hook, NY: Curran Associates. https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html.

[17] Sommer, Robin, and Vern Paxson. 2010. "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection." In 2010 IEEE Symposium on Security and Privacy, 305–316. Los Alamitos, CA: IEEE. https://doi.org/10.1109/SP.2010.25.

[18] Sowmya, T., and E. Anita. 2023. "A Comprehensive Review of AI Based Intrusion Detection System." Measurement: Sensors 28: 100827. https://doi.org/10.1016/j.measen.2023.100827.

[19] Yang, Zhen, Xiaodong Liu, Tong Li, Di Wu, Jingfeng Wang, Yunfeng Zhao, and Han Han. 2022. "A Systematic Literature Review of Methods and Datasets for Anomaly-Based Network Intrusion Detection." Computers & Security 116: 102675. https://doi.org/10.1016/j.cose.2022.102675.

[20] Al-Garadi, Mohammed Ali, Amr Mohamed, Abdulla Khalid Al-Ali, Xiaojiang Du, Ihsan Ali, and Mohsen Guizani. 2020. "A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security." *IEEE Communications Surveys & Tutorials* 22 (3): 1646–1685. https://doi.org/10.1109/COMST.2020.2988293.

[21] Axelsson, Stefan. 2000. "The Base-Rate Fallacy and the Difficulty of Intrusion Detection." *ACM Transactions on Information and System Security* 3 (3): 186–205. https://doi.org/10.1145/357830.357849.

[22] Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

[23] Berman, Daniel S., Anna L. Buczak, Jeffrey S. Chavis, and Cherita L. Corbett. 2019. "A Survey of Deep Learning Methods for Cyber Security." *Information* 10 (4): 122. https://doi.org/10.3390/info10040122.

[24] Ding, Derui, Qing-Long Han, Yang Xiang, Xiaohua Ge, and Xian-Ming Zhang. 2018. "A Survey on Security Control and Attack Detection for Industrial Cyber-Physical Systems." *Neurocomputing* 275: 1674–1683. https://doi.org/10.1016/j.neucom.2017.10.009.

[25] Ferrag, Mohamed Amine, Leandros Maglaras, Sotiris Moschoyiannis, and Helge Janicke. 2020. "Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study." *Journal of Information Security and Applications* 50: 102419. https://doi.org/10.1016/j.jisa.2019.102419.

[26] Giraldo, Jairo, Esha Sarkar, Alvaro A. Cárdenas, Michail Maniatakos, and Murat Kantarcioglu. 2017. "Security and Privacy in Cyber-Physical Systems: A Survey of Surveys." *IEEE Design & Test* 34 (4): 7–17. https://doi.org/10.1109/MDAT.2017.2709310.

[27] Humayed, Abdulmalik, Jingqiang Lin, Fengjun Li, and Bo Luo. 2017. "Cyber-Physical Systems Security—A Survey." *IEEE Internet of Things Journal* 4 (6): 1802–1831. https://doi.org/10.1109/JIOT.2017.2703172.

[28] Maglaras, Leandros A., Ki-Hyung Kim, Helge Janicke, Mohamed Amine Ferrag, Stylianos Rallis, Pavlina Fragkou, Athanasios Maglaras, and Tiago J. Cruz. 2018. "Cyber Security of Critical Infrastructures." *ICT Express* 4 (1): 42–45. https://doi.org/10.1016/j.icte.2018.02.001.

[29] Milenkoski, Aleksandar, Marco Vieira, Samuel Kounev, Alberto Avritzer, and Bryan D. Payne. 2015. "Evaluating Computer Intrusion Detection Systems: A Survey of Common Practices." *ACM Computing Surveys* 48 (1): Article 12, 1–41. https://doi.org/10.1145/2808691.

[30] Modi, Chirag, Dhiren Patel, Bhavesh Borisaniya, Hiren Patel, Avi Patel, and Muttukrishnan Rajarajan. 2013. "A Survey of Intrusion Detection Techniques in Cloud." *Journal of Network and Computer Applications* 36 (1): 42–57. https://doi.org/10.1016/j.jnca.2012.05.003.

[31] Moustafa, Nour, and Jill Slay. 2015. "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems." In *2015 Military Communications and Information Systems Conference*, 1–6. IEEE. https://doi.org/10.1109/MilCIS.2015.7348942.

[32] Pinto, Andrea, Luis-Carlos Herrera, Yezid Donoso, and Jairo A. Gutierrez. 2023. "Survey on Intrusion Detection Systems Based on Machine Learning Techniques for the Protection of Critical Infrastructure." *Sensors* 23 (5): 2415. https://doi.org/10.3390/s23052415.

[33] Sharafaldin, Iman, Arash Habibi Lashkari, and Ali A. Ghorbani. 2018. "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization." In *Proceedings of the Fourth International Conference on Information Systems Security and Privacy*, 108–116. SciTePress. https://doi.org/10.5220/0006639801080116.

[34] Tavallaee, Mahbod, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. 2009. "A Detailed Analysis of the KDD CUP 99 Data Set." In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 1–6. IEEE. https://doi.org/10.1109/CISDA.2009.5356528.