Research Square

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

Near-Deterministic Reconstruction of Enterprise Firewall Decisions for Policy Audit: Proxy Structure, Context Shift, and Residual Uncertainty

Özkan Canay

canay@sakarya.edu.tr

Sakarya University

Cem Özkurt

Sakarya University

Fatmir Garri

Sakarya University of Applied Sciences

Research Article

Keywords: Network security management, Firewall policy audit, Traffic decision reconstruction, Proxy variables, Class imbalance, Predictive uncertainty

Posted Date: August 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10586406/v1

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

Additional Declarations: No competing interests reported.

Near-Deterministic Reconstruction of Enterprise Firewall Decisions for Policy Audit: Proxy Structure, Context Shift, and Residual Uncertainty

Özkan Canay [1*], Cem Özkurt [2] and Fatmir Garri [3]

[1*]Department of Information Systems and Technologies, Sakarya University, Sakarya, Türkiye.
[2]Department of Data Science and Analytics, Sakarya University, Sakarya, Türkiye.
[3]Department of Information Technology, Sakarya University of Applied Sciences, Sakarya, Türkiye.

*Corresponding author(s). E-mail(s): canay@sakarya.edu.tr;
Contributing authors: cemozkurt@sakarya.edu.tr; fatmirgarri@subu.edu.tr;

Abstract

Near-perfect supervised performance on firewall logs can reflect deployed decision traces rather than independent threat detection. This study analyzes 1,048,576 Palo Alto traffic-log records with Allow, Drop, and Deny labels from a tool-capped 46-minute window. Direct label-source fields are excluded, and established tree ensembles serve as diagnostic instruments. A model-free audit finds seven minimum four-field determining sets in the all-field view, each including Source Port and session-end Bytes. Although 64.7 to 71.4% of rows occupy singleton key contexts, exact agreement persists in repeated contexts; train-only discovery retains exact seen-context agreement, but its test coverage is limited. Cardinality-preserving controls break exactness, and within a reduced 17-field view the best four-field subset reaches 99.9713% empirical agreement without exact determination. XGBoost and LightGBM reconstruct the proxy-rich core labels exactly, while removing volume and duration yields 0.9962 macro-F1. A stricter proxy-minimal view reaches 0.95 with XGBoost but is unstable with LightGBM. On identical application/category held-out rows, macro-F1 rises from 0.673 with ordinal LightGBM to 0.992 with native-categorical CatBoost. Temperature scaling changes fixed-threshold queue behavior materially across the two pipelines, making confidence-only review representation- and calibration-dependent. Marginal conformal summaries hide rare-Deny undercoverage in the random-split lineage, whereas an unscaled ordinal queue misses most context-shift errors. A public instantiation from the University of California, Irvine (UCI) Machine Learning Repository reproduces high reconstructability and rare-class collapse only for schema-independent components. These results support retrospective, claim-bounded decision reconstruction and simulated uncertainty diagnostics within one export, not attack detection, causal policy recovery, or cross-environment generalization.

Keywords: Network security management, Firewall policy audit, Traffic decision reconstruction, Proxy variables, Class imbalance, Predictive uncertainty

1 Introduction

Computer communication networks support critical services, so their security controls must remain auditable, reproducible, and resilient. Enterprise firewalls sit at this operational layer, observing large volumes of sessions, enriching them with application and policy context, and recording how the device handled each connection. These records are attractive for supervised learning because they already contain operational labels such as Allow, Drop, and Deny. They also have practical value, since a model

that reconstructs the decision behavior of a deployed firewall may support decision-consistency checks within policy-audit workflows, migration rehearsal, review of unexpected decisions, and reproducibility-oriented security analytics [1–4].

The same convenience introduces a methodological interpretation risk. A firewall decision is not equivalent to a human-validated intrusion label; it is a deterministic outcome produced by configured policy, application identification, zone and interface context, session termination behavior, and the logging pipeline. A model trained to predict Allow, Drop, or Deny may therefore learn the firewall's operational grammar, not an independent concept of maliciousness. This distinction matters most when the scores are perfect or nearly perfect, shifting the scientific question from which classifier performs best to why the target is so reconstructable from the remaining fields.

Work on leakage, shortcut learning, benchmark artifacts, dataset shift, and cross-dataset fragility in security machine learning reinforces this distinction [5–8]. Public intrusion-detection benchmarks often frame the target as benign versus attack, or as an attack family, whereas enterprise firewall labels are enforcement outcomes generated by the same system that writes many predictor fields. High performance may therefore be valid evidence of a stable decision surface inside one export while remaining weak evidence for general attack detection.

The empirical setting is a real enterprise Palo Alto firewall traffic-log export at a million-record scale. Its tool-capped file contains a 46-minute and 34-second observed timestamp span, not a designed monitoring period, and supports a three-class operational decision target of Allow, Drop, or Deny. Because the class distribution is strongly imbalanced, with Allow dominant, Drop substantial, and Deny rare, accuracy alone is insufficient. The evaluation therefore prioritizes balanced accuracy, macro-F1, and per-class confusion matrices alongside time-ordered split checks, cross-validation, and uncertainty diagnostics. This places the study at the intersection of operational decision reconstruction, approximate reasoning over policy-conditioned labels, and reproducible security analytics. The zero-conflict audit establishes no empirical label conflict for repeated contexts observed in this export. Uncertainty revealed by proxy reduction and context shift is therefore consistent with evidence-relative epistemic uncertainty, while unobserved contexts or the wider deployment population could still contain aleatoric conflict [9].

Beyond methodological rigor, this framing holds practical relevance for network and systems management. In a managed communication environment, firewall logs are more than security artifacts; they provide empirical measurements of how an enforcement point operationalizes configured policy for observed traffic [10]. A reconstruction model can therefore be read as an empirical model of decision behavior, while uncertainty diagnostics identify the contexts where that behavior becomes less stable, less confidently reconstructed, or more dependent on proxy structure. Deterministic policy analysis already covers part of this ground. Rule-level techniques detect shadowing, redundancy, and related configuration anomalies directly from the rule set [11], and anomaly-classification tools rank those findings so that security managers can act on them [12]. Those methods read the configured policy; the present study reads the decisions that policy produced, so the two views answer different audit questions. Recent learning work on campus-network firewall logs targets anomalous records [13], while earlier firewall-action studies report high learnability without asking why the label is reconstructable.

The experimental design separates direct reconstruction, proxy-reducing ablations, context-held-out stress tests, and uncertainty-aware diagnostics. The core setting excludes direct label-source fields and forms the main claim, while the with-policy setting is reported strictly as a diagnostic upper-bound condition, not as the primary result. Feature-family ablations test whether near-deterministic reconstruction collapses when application, volume, transport, or zone-related context is unavailable. Context-held-out tests then ask whether the learned decision surface remains stable for unseen application/category, destination-service, or rule contexts. This design cannot remove every indirect proxy, but it makes proxy structure and residual uncertainty visible instead of hiding them behind a single score table.

Establishing the claim boundary forms a core contribution of this work; the study does not seek to determine the inherent maliciousness of Drop and Deny records, nor the benignity of Allow records. Instead, it investigates how far a deployed firewall's logged decision surface can be reconstructed from the remaining structured fields, how much uncertainty remains after proxy reduction and context shift, and what near-perfect reconstruction implies for evaluation practice in cybersecurity machine learning. This makes the study closer to uncertainty-aware operational decision reconstruction and policy-consistency analysis than to conventional intrusion detection.

The investigation is guided by three research questions. *RQ1* asks how far scalable tabular classifiers can reconstruct enterprise firewall decisions (Allow, Drop, Deny) from structured log context after direct label-source fields are excluded, and how much this reconstruction depends on prominent application-layer proxies. *RQ2* examines whether the apparent decision boundary remains stable under temporal and

context-held-out stress tests, where familiar operational combinations are no longer guaranteed. *RQ3* asks whether calibration, entropy, conformal prediction sets, and selective classification can separate near-deterministic regions from residual ambiguity in a way that is useful for operational review rather than attack-detection claims.

Together, these questions support one claim. A near-perfect firewall-decision score is best read as a measurable boundary between deterministic and approximate regions of a policy-conditioned label. The paper develops this claim through four contributions:

- A model-free reconstructability audit shows exact context-level determinism in the full export and exactly seven four-field minimum determining keys, with explicit repeated-row support and held-out context coverage.
- Boundary-stress diagnostics combine temporal and feature-family tests with an out-of-vocabulary (OOV)-aware, paired model and encoding sensitivity analysis for context-held-out shift.
- Conformal and selective diagnostics expose rare-Deny undercoverage, Mondrian conformal coverage recovery within the exchangeable regime, and high-confidence errors that a confidence-only queue fails to capture.
- A bounded external instantiation on an independent public firewall dataset reproduces high reconstructability, redundant duplicate structure, and rare-class collapse for the evaluated schema-independent components, while the policy-conditioned proxy layer remains specific to the enterprise export.

The remainder of this paper is organized as follows. Section 2 reviews firewall-decision learning, security-ML evaluation risks, proxy structure, and uncertainty-aware prediction. Section 3 defines the traffic-decision target, feature views, determinism audit, and uncertainty diagnostics. Section 4 describes the evaluation protocol, conformal and selective-classification settings, and reproducibility controls. Section 5 reports reconstruction, ablation, context-shift, calibration, conformal, and review-queue results. Section 6 discusses the operational interpretation and limitations, and Section 7 concludes with the implications for firewall-log analytics.

2 Related Work

The relevant literature spans direct firewall-decision modeling, scalable tabular learning, class-imbalance evaluation, leakage-aware security ML, and uncertainty-aware prediction.

2.1 Machine Learning for Firewall and Network Security Logs

Machine learning has already been applied to firewall decision and action classification, but most direct studies use much smaller datasets than the enterprise traffic-log corpus studied here. Al-Haija and Ishtaiwi modeled firewall decisions on the Internet Firewall-2019 dataset and reported high performance with an optimizable decision tree and a shallow neural network [1]. Rahman et al. later studied multiclass classification of internet firewall log files into action categories such as Accept, Drop, Deny, and Reset-Both, with Random Forest reported as a strong model [14]. Al-Behadili also used decision-tree learning for multiclass firewall access classification [15], while Aljabri et al. compared several classical and neural models for firewall log action classification and reported Random Forest as the best-performing approach [2]. More recently, Efeoglu and Tuna analyzed real firewall logs, demonstrating that tree algorithms can achieve near-perfect multi-action classification [16]. Similarly, Fatima et al. introduced an explainable AI-driven firewall evaluation framework, emphasizing that tree-based models such as XGBoost provide both high accuracy and critical interpretability for optimal network defense [17].

These studies establish that firewall actions are highly learnable from structured fields, but they seldom distinguish generalized security behavior from deployed policy structure and its operational proxies. Firewall decisions are device-generated outcomes, not independent labels created by human analysts. Recent firewall-related work also addresses class imbalance, error resilience, formal verification, packet filtering, deep learning, and ensemble tuning [4, 18, 19]. Korkmaz et al. used SMOTE variants for underrepresented firewall-action classes [19], while Bairwa et al. compared tree and boosting models for TCP firewall action prediction [18]. This literature confirms the tractability of firewall action modeling but leaves near-perfect prediction in large enterprise logs methodologically underexamined.

A complementary and older line of network and systems management research analyzes firewall policies directly instead of learning from logs. Al-Shaer and Hamed modeled and managed firewall policies and formalized the discovery of intra- and inter-firewall rule anomalies [20]; Yuan et al. built FIREMAN to detect rule conflicts, shadowing, and inconsistencies through static analysis of the rule base

[21]; and Hu et al. detected and resolved firewall policy anomalies through a rule-based segmentation of the policy space [22]. This tradition operates on the configured rule set with white-box access to policy semantics and yields exact, causal statements about rule ordering, coverage, and conflict. The present study is deliberately complementary; it works from the logged decision outcomes without access to the rule engine, so it reconstructs the decision surface as an empirical, black-box witness and does not verify the policy itself. The two views are mutually reinforcing (rule-graph analysis explains why a decision should occur, whereas log-based reconstruction and its uncertainty diagnostics flag where observed decisions leave the near-deterministic core and warrant policy review), but they must not be conflated, because a reconstruction model cannot substitute for deterministic rule analysis or policy ownership.

The determining-key audit also has a methodological antecedent in database profiling. TANE searches efficiently for exact and approximate functional dependencies by partitioning rows [23], while HyFD combines focused sampling with full-data validation to discover minimal functional dependencies at scale [24]. The present work does not propose a new dependency-discovery algorithm. Its exhaustive census is deliberately narrower, testing whether subsets of at most four audited log fields determine one operational outcome and then examining the support and held-out coverage of the resulting keys.

2.2 Scalable Tree-Based Models for Categorical Security Telemetry

Tree-based models remain common in network security and traffic-classification studies because they handle mixed tabular features, nonlinear interactions, and categorical encodings well. In the broader intrusion detection literature, Random Forest, Extra Trees, XGBoost, LightGBM, and CatBoost are frequently used either as primary classifiers or as components in stacked ensembles [25, 26]. Talukder et al. used tree-based models with oversampling and feature embedding for large and imbalanced intrusion detection datasets [25]. Other studies have combined ensemble learning, deep models, and traffic-specific architectures to improve performance on benchmark traffic datasets [27, 28]. Despite the rising popularity of deep models, Zoppi et al. systematically compared classifiers across numerous tabular intrusion datasets and found that no deep neural network consistently outperforms tree-based classifiers [29]. Alam et al. also demonstrated that tree ensembles paired with SHAP can provide interpretable anomaly detection directly from system logs [30].

Although this work supports tree ensembles for enterprise telemetry, the present target differs from normal-versus-attack labels or attack families; it is the operational Allow, Drop, or Deny decision. Strong performance may therefore reconstruct deterministic device behavior instead of recognizing an attack. Model choice and runtime are reported, but interpretation centers on the label-generating process. Explainable tree-based intrusion detection studies, such as the SHAP-based ensemble-tree work by Le et al. [26] and recent XAI-oriented IDS work using SHAP, LIME, or ELI5 [28, 30, 31], also provide useful context. They show how feature attribution can help inspect model behavior in security telemetry. The present study does not use XAI to make causal claims, but this literature supports the broader need to inspect which feature groups drive high-performing security-log models.

Recent survey and deployment-oriented work also shows that model comparisons in network intrusion detection have shifted from isolated classifier accuracy toward data quality, scalability, adversarial resilience, privacy, interpretability, deployment constraints, and enterprise log-analytics integration [31–34]. Anis et al., for example, review deep-learning NIDS studies from 2020 to 2025 and identify several of these recurring limitations [32]. Such surveys and deployment-oriented studies are useful background for the present study, but they are not direct evidence about firewall decision labels. The direct evidence still comes from firewall-action studies, while the survey literature helps frame why a high-performing log model must be evaluated as part of an operational data pipeline.

2.3 Class Imbalance and Evaluation Metrics

Class imbalance is a recurring problem in network security datasets. In intrusion detection benchmarks, majority classes can dominate global metrics, while rare attack classes remain poorly detected [35, 36]. Bulavas et al. studied highly imbalanced multiclass network intrusion datasets and used metrics such as balanced accuracy to expose model behavior beyond raw accuracy [35]. Bagui and Li discussed resampling for imbalanced intrusion detection datasets [36]. Recent large-scale IDS studies continue to report very high scores, but these results often depend on balancing, feature selection, synthetic data generation, or benchmark-specific preprocessing [25, 37–39].

The same metric problem appears in firewall decision modeling. In the present dataset, Allow represents most records, Drop is much smaller, and Deny is rare. Accuracy is therefore reported for comparability, but it is not sufficient. Macro-F1, balanced accuracy, and per-class scores are necessary

to show whether the minority decision classes are also reconstructed [35, 36]. The Results section therefore reports macro-F1 and balanced accuracy next to accuracy, and the no-application-context ablation includes per-class scores.

2.4 Label Leakage, Proxy Variables, and Dataset Bias

A central methodological challenge in this context is how to interpret near-perfect classification scores. Prior cross-domain analyses show that data leakage and reproducibility failures can inflate reported performance metrics [5]. Apicella et al. further classify leakage risks across machine-learning and transfer-learning pipelines [40]. Although these papers are not firewall-specific, their warning applies directly to operational security logs because a feature can be technically different from the target and still act as a proxy for it.

Cybersecurity-specific surveys and cross-evaluation studies point to similar concerns. Chou and Jiang review data-driven network intrusion detection and discuss the limits of benchmark datasets, simulated traffic, imbalance, and generalization [41]. Apruzzese et al. evaluate ML-based network intrusion detection across datasets and show why single-dataset results can fail under dataset shift and changing attack distributions [6]. Maseer et al. also highlight benchmarking weaknesses in CICIDS2017-style evaluations [37]. These studies do not directly model Allow, Drop, and Deny labels, but they support the caution that high performance in security ML must be interpreted through data provenance, label construction, and evaluation design.

In enterprise firewall telemetry, proxy risk is high. Operational labels are deterministically generated by device policies and recorded alongside network features shaped by those exact same policies. Excluding direct label-source fields is necessary, but it cannot guarantee that all proxies have been removed. This paper therefore uses feature-set variants and ablations not to claim perfect leakage removal, but to make the proxy structure explicitly visible.

Recent IDS work makes this concern concrete. Leakage-free protocols, multi-dataset evaluations, and cross-dataset transformers show that strong within-dataset scores can reflect artifacts or split design instead of generalizable security reasoning [42–44]. Lifecycle-aware datasets, shortcut-detection methods, and protocol-specific traffic studies expose related dependence on data provenance and operational regularities [7, 8, 45]. The parallel for the present study is that an Allow label may describe many operational contexts, and a model can learn those contexts as proxies for the firewall decision.

2.5 Uncertainty Quantification, Conformal Prediction, and Selective Classification

The diagnostic layer draws on conformal prediction and adaptive prediction sets for set-valued reconstruction [46–48]. Within approximate reasoning, nested conformal sets can be read as consonant-plausibility level sets [49], while entropy and confidence are task-dependent uncertainty summaries [50]. Selective classification supplies the operational abstention layer [51], and Brier score and expected calibration error (ECE) diagnostics quantify probability quality. These instruments are used here to locate where a policy-conditioned operational label is deterministic, ambiguous, or class-conditionally undercovered.

This positioning connects the study to a broader approximate-reasoning question about the source and structure of predictive uncertainty. The distinction between aleatoric uncertainty, which is irreducible noise in the data-generating process, and epistemic uncertainty, which reflects a learner's limited evidence and is in principle reducible, is central to modern uncertainty quantification [52]. A firewall decision target is unusual in this respect; the model-free audit finds no empirical label conflict among repeated contexts in this export, and proxy reduction and context shift expose evidence-relative uncertainty consistent with an epistemic interpretation. This conclusion is limited to the observed export; it does not rule out aleatoric conflict in unobserved contexts or in the wider deployment population.

The validity of conformal prediction, in turn, rests on exchangeability [46–48], and marginal coverage averaged over a heavily imbalanced label can silently underserve a rare class; class-conditional (Mondrian) conformal prediction and set-valued classifiers with bounded per-class error improve per-class coverage within the exchangeable regime [46, 53]. What remains comparatively unexamined, and what this study contributes on the approximate-reasoning side, is how these validity and imprecision notions behave at the extreme of a near-deterministic, policy-conditioned, and severely imbalanced operational label, where marginal validity is most apt to mislead and where the boundary between determinism and residual ambiguity can be measured directly instead of assumed.

Table 1 Selected recent WoS-indexed journal studies used to position the present protocol.

Study group	Firewall or security task	What prior work already covers	Role in this paper
Leakage quantification [54]	Information leakage from observable system variables.	Bayes-optimal log loss and accuracy were connected to mutual-information estimation and tested on synthetic and OpenSSL TLS data.	Provides a methodological analogue for testing whether near-perfect prediction reflects revealing observables.
Drift-aware intrusion detection [55]	Network intrusion detection under concept drift.	Self-supervised representation enhancement and weakly supervised tuning were evaluated offline and online under varying drift.	Motivates explicit separation of stationary-export evidence from adaptation claims.
Probability calibration [56]	Calibration of Random Forest probability estimates.	Multiple post-hoc methods and tuned forests were compared on synthetic and 30 real datasets using ECE, Brier score, and log loss.	Supports treating confidence and fixed-threshold review queues as calibration-dependent.
Cross-dataset transfer [57]	Federated intrusion detection evaluated across IoT datasets.	Dataset-centric evaluation showed substantial degradation when models were moved outside their training datasets.	Bounds claims drawn from a single enterprise export and motivates the context-held-out tests.
Cross-domain recalibration [58]	Deep traffic classifiers across virtual and physical network-function domains.	Cross-domain generalization was analyzed statistically and required recalibration outside the training domain.	Supports reporting calibration and conformal behavior separately from raw accuracy under shift.
Policy anomaly triage [12]	Anomaly classification and resolution priority in firewall policy configuration.	An anomaly classification model and a visual decision-support tool help security managers detect and prioritize firewall policy anomalies.	Establishes the policy-management use case this reconstruction serves, without a learned decision surface or uncertainty reporting.
Firewall log learning [13]	Anomaly detection from campus-network firewall logs.	A hybrid graph convolutional and graph attention model was trained on operational firewall log records.	Closest prior work on learning from firewall logs; it targets anomalous records rather than the reconstructability of the enforcement decision.
Logic-based policy analysis [11]	Shadowing, redundancy, correlation, generalization, and irrelevance anomalies in firewall rule sets.	First-order logic rules with a knowledge-graph view detect policy anomalies without training data and with formal correctness.	Marks the deterministic alternative this study complements; determining-key evidence here is empirical and export-specific, not a formal proof.
Evaluation realism [59]	Label timing in malware detection pipelines.	Label delays were shown to change measured detection performance relative to conventional offline evaluation.	Supports treating post-decisional field availability as an evaluation-design variable rather than a fixed benchmark property.
Network-management log analytics [60]	Abnormal network switch states from log and NETCONF state data.	Log rarity statistics combined with resource state reached an F1 score of 0.86 on the authors' collected dataset.	Places firewall-decision reconstruction within operational network-management analytics rather than intrusion-detection benchmarking.

Notes. The selection is restricted to published journal articles in the Web of Science Core Collection (SCIE): seven from 2025 and three from 2026. The present protocol combines label-source exclusion, proxy controls, context-held-out tests, determining-key audits, and uncertainty diagnostics on one enterprise export.

2.6 Recent Firewall and Network-Security ML Studies

Recent work on leakage quantification, drift adaptation, probability calibration, and cross-environment evaluation is summarized in Table 1. The narrower contribution here is their joint use with determining-key and proxy evidence to explain one enterprise export.

3 Methodology

3.1 Study Workflow and Claim Boundary

The workflow begins with a real enterprise Palo Alto traffic-log export and proceeds through target construction, label-source exclusion, proxy-control feature views, direct determinism auditing, and uncertainty-aware evaluation under a common preprocessing pipeline. The common protocol and its three evidence lanes are summarized in Figure 1, which shows how they support the claim-bounded operational interpretation.

Claim boundaries are set by the nature of the target and the meaning of the ablations. The target is an operational firewall decision, not an analyst-confirmed threat state, and removing application and category context is interpreted as a stress test, as opposed to proof that every indirect proxy has been eliminated. Each observation is one structured traffic-log record; no session sequences, user histories, rule graphs, or packet payloads are used. The reported outcomes therefore measure policy-conditioned decision reconstruction rather than temporal threat behavior.

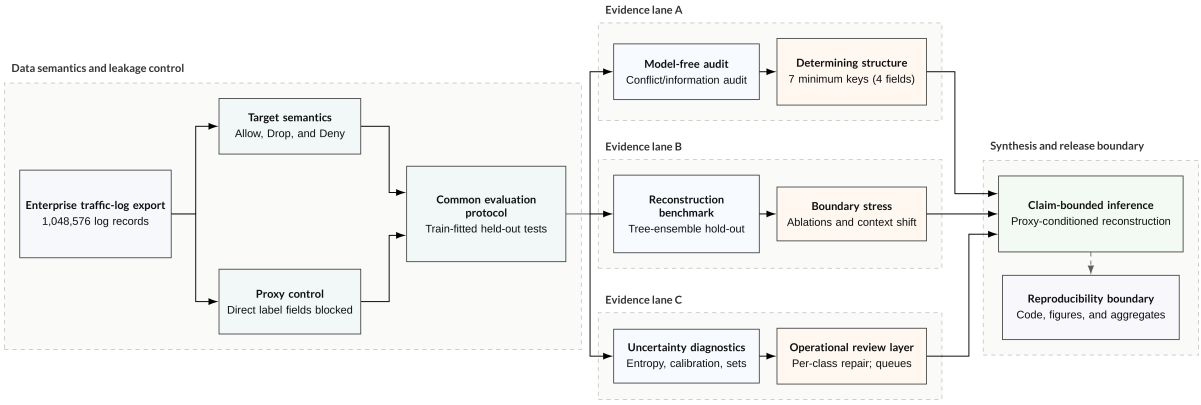


Fig. 1 Evidence architecture for uncertainty-aware firewall decision reconstruction

3.2 Data Source and Confidentiality

The dataset consists of 1,048,576 Palo Alto firewall `TRAFFIC` log records from a real enterprise next-generation firewall deployment. The CSV contains one header row and exactly 2^{20} data rows, so the export boundary is tool-defined rather than an operationally chosen duration. The raw rows are monotonically ordered from newer to older timestamps, and time-dependent analyses sort them into ascending order. The observed timestamp span is about 46 minutes and 34 seconds at roughly 375 records s^{-1}; the available artifacts do not establish how much older traffic was omitted or whether this interval represents other periods. The preprocessing script reads the entire available CSV in chunks without an additional row cap. Raw and record-level processed logs are not redistributed because timestamps, application signatures, zones, interfaces, port configurations, and firewall rules can expose sensitive operational structure. The processed analysis dataset contains no source or destination IP addresses and no user identifiers, so these record-level identifiers were unavailable to all models. The public package provides code, aggregate outputs, diagnostic summaries, figures, and schema-level documentation; full reruns require authorized local access to the institutional data or an equivalently approved dataset.

A capped sample of 113,809 records, built with a 50,000-record per-class cap that retains all 13,809 Deny records and yields a rebalanced 43.9/43.9/12.1 class mix, was also used for faster cross-validation experiments. The capped sample is not the main evidence for the paper; because its class distribution differs from the full export, it is used only to test whether the near-perfect pattern and the model ranking persist under repeated stratified evaluation at lower computational cost, not for absolute-level comparison with the full-data metrics. The full-data hold-out experiments remain the primary basis for the claims.

3.3 Target Construction and Class Imbalance

The three operational decisions (Allow, Drop, and Deny) are not independently adjudicated benign or malicious states. Allow constitutes 76.73% of the records (804,586), Drop accounts for 21.95% (230,181), and Deny represents 1.32% (13,809). Raw accuracy is documented for basic comparability but is not sufficient for evaluating model fidelity. Macro-F1, balanced accuracy, and per-class scores are prioritized so that the minority Deny class is not obscured by the majority Allow class.

Target construction takes the traffic subtype (`Threat/Content Type`) as its source field, with action and session-end reason used as corroborating checks. Every observed source-field tuple used to construct the target is enumerated in Table 2. Traffic subtype has precedence because all Drop records also end in policy-deny, and action alone cannot separate the classes because deny-subtype sessions occur under both allow and drop. Here, decision denotes this three-class composite traffic-log endpoint, not the raw action field, an attack label, or a complete account of every enforcement layer.

The Deny class is operationally heterogeneous, with 9,062 records ending in policy-deny, 3,621 (26.2%) authentication-policy redirects logged with action allow, 1,121 records lacking a recorded end reason, and 5 records ending in auth-policy-deny. Redirected sessions are retained in Deny because the requested session was not permitted to proceed as requested. This labeling choice does not drive the results; excluding the 3,621 auth-policy-redirect records from Deny and re-running the full-data core hold-out leaves reconstruction essentially unchanged (XGBoost still makes zero errors at macro-F1 1.000000, and LightGBM makes three errors at macro-F1 0.999752, against zero at baseline), so near-perfect reconstruction is insensitive to how redirected sessions are labeled. Allow refers strictly to the

Table 2 Observed source-field combinations used to construct the three-class target.

Raw action	Traffic subtype	Session-end reason	Target	Records
allow	end	aged-out	Allow	537,905
drop	drop	policy-deny	Drop	230,181
allow	end	tcp-fin	Allow	152,929
allow	end	tcp-rst-from-client	Allow	77,565
allow	end	tcp-rst-from-server	Allow	26,525
allow	end	threat	Allow	9,658
drop	deny	policy-deny	Deny	8,727
allow	deny	auth-policy-redirect	Deny	3,621
drop	deny	missing	Deny	1,121
allow	deny	policy-deny	Deny	335
allow	deny	auth-policy-deny	Deny	5
allow	start	missing	Allow	3
allow	end	tcp-reuse	Allow	1

Table 3 Operational field lineage and analytical use.

Field timing status	Representative fields	Timing and analytical use
Direct label source	Raw action, traffic subtype, session-end reason	Target construction only; excluded from every predictor view
Decision-time availability unverified	Zones, interfaces, protocol, ports, countries	Included in core and proxy-minimal stress views; exact decision-time availability was not observed
Session-end application/enrichment	Application taxonomy, risk, SaaS, AI-traffic flags	Included in core; removed as a family in the no-application-context ablation
Post-decision telemetry	Bytes, packets, elapsed time	Included in core; removed in the reduced-telemetry view because these fields are post-decisional
Policy-proximal	Rule, Action Source	Excluded from core; added only in the with-policy upper-bound condition

Notes. Timing follows log semantics rather than device-internal observation. No feature view is presented as a verified pre-decision representation.

traffic-policy decision; 9,658 Allow sessions were subsequently terminated by threat enforcement (end reason threat), which belongs to a separate enforcement layer outside this target.

Fields used directly in target construction were excluded from the core feature set. This exclusion is necessary but does not remove all proxy risk, since other operational fields can encode the same policy decision indirectly; the paper therefore reports feature-set variants instead of a single headline benchmark.

3.4 Feature Configurations and Proxy-Control Design

The full-data experiments use three feature settings that differ in proximity to policy and application context. The core baseline contains 24 traffic, application, zone, port, country, byte, packet, elapsed-time, risk, SaaS, and AI-traffic fields after excluding the direct label-source columns. The with-policy setting adds two policy-related fields and is used only as an upper-bound diagnostic. The no-application-context ablation removes Application, Category, Category of app, Subcategory of app, and Technology of app, leaving 19 features to test whether reconstruction remains stable when the most explicit application-layer proxy family is suppressed.

This no-application-context ablation is the primary stress test for whether reconstruction depends on application identification or categorical metadata. It is a proxy-control experiment, not a guarantee of proxy-free inference, because zones, interfaces, ports, and byte counts are also shaped by session handling policies. Its value is comparative, measuring whether the decision boundary remains stable when an overt proxy family is suppressed.

One further distinction matters for interpretation. Zone, interface, transport, and country fields are treated as session-context candidates, whereas application fields in session-end logs can reflect the final App-ID after application identification completes [61]. Volume and duration are recorded at session end and are partly mechanical consequences of the decision itself. In this export, every Drop record has zero received bytes and zero elapsed time, and 96.6% of Drop records carry the application placeholder not-applicable because the session ended before application identification completed. The resulting analytical roles are made explicit in Table 3.

The without-volume/duration view reported in Section 5.7 removes major post-outcome fields but retains final application and related metadata; it is therefore treated as a reduced telemetry view rather

than a verified decision-time feature set. A stricter proxy-minimal stress view, reported in the same section, retains only zone/interface and transport/port fields, with a country-context variant, and is used as a conservative lower-bound check, not as a causal or proxy-free representation.

Six operational feature families are retained in the core setting so that the proxy structure remaining after direct label-source fields are removed is explicit. These families are application context (Application, Category, Subcategory of app, Category of app, Technology of app), the major proxy family tested by the no-application-context ablation; zone and interface context (Source/Destination Zone, Inbound/Outbound Interface); transport and ports (IP Protocol, Source/Destination Port); country context (Source/Destination Country); volume and duration (Bytes, Bytes Sent/Received, Packets, Packets Sent/Received, Elapsed Time), captured at session end and partly post-decisional; and vendor-enriched risk, SaaS, and AI-traffic flags. Direct label-source fields (raw action, raw traffic subtype, raw session-end reason) are used only to construct the target and excluded from every view, bookkeeping and timestamp fields (Receive Time, Generate Time, High Res Timestamp, Type) are dropped as constant or time-memorization risks, and policy-proximal fields (Rule, Action Source) are added only in the with-policy upper-bound setting.

The later LightGBM diagnostic pass extends this design because LightGBM reached the full-data ceiling and had the highest mean macro-F1 in capped-sample cross-validation. It adds the time-ordered split, leave-one-feature-family-out variants, strict feature restrictions, and context-held-out tests used to locate the operational boundary of reconstruction; the exact split sizes and metrics are reported with the corresponding Results tables.

3.5 Model-Free Determinism and Information Audit

Independently of any trained classifier, the reconstructability of the target is measured directly from the data. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the $N = 1{,}048{,}576$ exported records, $Y \in C$ the decision label with class set C, and V the tuple of fields in a given feature view, taking values v in the observed context set $\mathcal{V}$. Writing $N_v = |\{i : V(x_i) = v\}|$ for the number of records in context v and $N_{v,c} = |\{i : V(x_i) = v,\, y_i = c\}|$ for those also labeled c, the empirical conditional label distribution in context v is $\hat{p}(c \mid v) = N_{v,c}/N_v$. For every feature view, records are grouped by their exact value combination, and the reported split-independent quantities include the number of distinct contexts $|\mathcal{V}|$; the number of conflict rows (records whose context carries more than one decision label),

$$\mathrm{Conf}(V) = \sum_{v \in \mathcal{V}} N_v \, \mathbf{1}\{\max_{c \in C} \hat{p}(c \mid v) < 1\}\,; \tag{1}$$

the empirical conditional entropy of the decision given the view, in bits,

$$H(Y \mid V) = -\sum_{v \in \mathcal{V}} \frac{N_v}{N} \sum_{c \in C} \hat{p}(c \mid v)\, \log_2 \hat{p}(c \mid v), \tag{2}$$

with the convention $0 \log_2 0 = 0$; and the in-sample resubstitution error (the empirical majority-vote error) obtained by assigning each context its majority label,

$$\mathrm{Err}^{\star}(V) = \sum_{v \in \mathcal{V}} N_v \left(1 - \max_{c \in C} \hat{p}(c \mid v)\right). \tag{3}$$

In Eqs. (1)–(3), $\mathbf{1}\{\cdot\}$ is the indicator function. These quantities are properties of the export itself and assume nothing about models; zero conflict rows, equivalently $H(Y \mid V) = 0$ and $\mathrm{Err}^{\star}(V) = 0$, certify only empirical determination on the analyzed records. Because high-cardinality views leave many contexts unique, the audit also reports the unique-column-combination proximity $|\mathcal{V}|/N$, singleton-context and singleton-row shares, and the row mass in contexts with support of at least 2, 5, and 10. These are descriptive properties under the export's empirical distribution, not estimates of a population Bayes error.

The audit then searches directly for minimum determining sets by exact grouping over field subsets. Writing V_S for the sub-tuple of fields restricted to a field subset S drawn from the 24 audited fields, S is a *determining set* of the decision label when it admits zero conflict rows, equivalently when the conditional entropy of Eq. (2) restricted to V_S vanishes:

$$\mathrm{Conf}(V_S) = 0 \iff H(Y \mid V_S) = 0, \tag{4}$$

so that Y is an exact function of V_S on the export. Greedy forward selection, which adds the field that most reduces $\mathrm{Err}^\star$ of Eq. (3) until it reaches zero, serves for discovery and is completed by an exhaustive census. Every subset of at most three fields (2,324 candidates) and every four-field subset (10,626 candidates) is checked against Eq. (4) by exact grouping, with each check reduced to a single integer-key sort. The full-export census is retained as a structure audit. A second census repeats the same exhaustive search using only the seed-42 stratified training partition; candidate field sets are then frozen. Their majority-label mappings are fitted in the corresponding stratified training data and, as a temporal stress check, in the earliest 80% of records before evaluation on the untouched stratified or latest-20% test block. This separates candidate discovery from stratified test evaluation and tests the frozen field sets under chronological extrapolation.

Two controls probe whether exactness is a high-cardinality artifact. First, Source Port is reassigned across records in 20 prespecified permutations, preserving its marginal frequencies and cardinality while breaking its joint association with the other key fields and the label. Second, Source Port is summarized by class-conditioned well-known, registered, and dynamic/private port bands and by its stand-alone majority-label error. The reduced 17-field telemetry view is searched for the lowest empirical majority-label error through four fields even when no exact key exists. These checks distinguish an export-specific fingerprint from a candidate policy rule.

The train-fitted exact-match lookup, defined as a majority label per seen context with a global-majority fallback, acts as a model-free memorization witness. On seen test contexts, it incurs only within-context conflict loss, being exact in conflict-free views and tracking the scaled in-sample floor in conflicted views up to majority ties and split composition, so its unseen-context failures mark the genuine generalization burden. Model success then certifies that the information limit remains attainable for learners that must generalize; a model shortfall is not evidence of unattainability, since it may reflect objective choices such as balanced class weighting (Section 5.2). These split-independent quantities and the two witnesses form the reconstructability audit against which the trained classifiers are read.

3.6 Preprocessing and Models

The initial capped-sample screening experiments covered six supervised tabular classifiers, namely Decision Tree, Random Forest, Extra Trees, XGBoost, LightGBM, and CatBoost. The full-data analysis then focused on four scalable tree-based models that were most relevant for large structured logs and near-perfect sample behavior. Extra Trees was chosen as a strong ensemble baseline for tabular data with mixed feature types. XGBoost was included as a widely used gradient boosting model for structured classification tasks. LightGBM was selected for its high efficiency on large tabular datasets, and CatBoost was integrated because it is designed to handle categorical structure effectively. The three gradient-boosting frameworks follow a regularized additive-tree paradigm; LightGBM differs mainly through histogram-based, leaf-wise tree growth, and CatBoost through ordered boosting and optional native handling of categorical features. Extra Trees instead aggregates trees built with extremely randomized split selection, which decorrelates ensemble members. Because these are established methods used here as diagnostic instruments, their formal objectives are not restated.

Categorical fields were imputed and ordinally encoded inside a reproducible machine-learning pipeline, with unseen categories mapped to a reserved value. This common ordinal representation was used by all four published full-data ensemble benchmarks, including CatBoost; native categorical handling is evaluated only in the review-motivated sensitivity analysis described below. Numeric fields were median-imputed, and scaling was applied only for models that required it during the wider sample-screening stage. The preprocessing pipeline was fitted only on the training portion of each hold-out split and inside each cross-validation fold, which avoids fitting encoders or imputers on held-out labels.

The model settings were fixed before the final full-data experiments rather than tuned against the hold-out test set. Extra Trees used 200 estimators with balanced class weights. XGBoost used 300 trees, depth 6, learning rate 0.08, subsampling 0.9, column subsampling 0.9, and the histogram tree method. LightGBM used 300 estimators, learning rate 0.08, and balanced class weights. CatBoost used 300 iterations, depth 6, learning rate 0.08, multiclass loss, and automatic balanced class weights. This design keeps the comparison reproducible and avoids turning the paper into a hyperparameter search exercise. Because the selected models are expected to sit near the numerical ceiling on this deterministic target, pairwise significance testing between them is intentionally avoided as uninformative. Deep tabular baselines are left for future work because the present paper focuses on scalable tree ensembles and on the methodological meaning of near-deterministic operational labels.

Table 4 Experimental lineages and their comparison boundaries.

Lineage	Partition and scope	Models and weighting	Permitted comparison
Published ensemble benchmark	Stratified 80/20 core, with-policy, and no-application views	Four ensembles with documented model-specific weighting	Within-lineage ceiling, errors, and runtime; no algorithm-superiority claim
LightGBM diagnostics	Stratified, time-ordered, feature-restricted, and five-seed group partitions	Balanced-weight LightGBM with ordinal encoding	Boundary sensitivity within one fixed diagnostic pipeline
Conformal and selective	60/20/20 train, calibration, and test lineage	Retrained balanced-weight LightGBM	Coverage, set structure, and retained-risk diagnostics
Review sensitivity	Paired seed-42 context partitions and shallow core trees	Four model/encoding choices plus fixed-depth CART	Representation and capacity sensitivity on identical rows

Notes. Results are compared directly only within a row. Cross-lineage differences can also reflect split, calibration, weighting, encoding, or software-environment changes.

A separate interpretability sensitivity fits single decision trees at four depths fixed in advance (4, 6, 8, and 12) on the full-data core 80/20 split. Every prespecified depth is retained, so no test-selected tree is promoted as a tuned baseline.

4 Experimental Design

The study protocol maps each research question to a defined diagnostic layer: direct hold-out reconstruction and proxy-reducing ablations address *RQ1*; within-window and context-held-out checks address *RQ2*; and calibration, conformal, and selective diagnostics address *RQ3*.

4.1 Evaluation Protocol

The final experiment set instantiates the feature views and diagnostics defined above through full-data core, with-policy, and no-application-context hold-outs; capped-sample cross-validation; a LightGBM temporal and feature-family diagnostic pass; uncertainty, conformal-prediction, and selective-classification diagnostics; a model-free determinism audit covering duplicate structure, exact-match lookup, view-conditional information limits, and minimum determining keys; determining-key effective-support checks; and review-motivated model, encoding, OOV, and shallow-tree sensitivity analyses.

For the main hold-out analyses, the full dataset was split into 838,860 training records and 209,716 hold-out test records using a stratified 80/20 split with random state 42 across full-data settings. The capped sample was used for five-fold stratified cross-validation to test whether model ranking and near-perfect behavior persist under repeated evaluation at lower computational cost.

Reported metrics include accuracy, balanced accuracy, weighted precision, weighted recall, weighted F1, macro-F1, fit time, prediction time, total model time, and prediction milliseconds per 1,000 rows. Macro-F1 and balanced accuracy are primary because they are less dominated by the majority Allow class than accuracy; per-class scores and confusion patterns are reported where aggregate scores could hide Deny behavior. For the temporal diagnostic, the earliest 80% of records in timestamp order were used for training and the latest 20% for testing; the test set contained 156,871 Allow, 50,309 Drop, and 2,536 Deny records.

The paired context sensitivity preserves the ascending timestamp order and seed-42 group partitions of the published diagnostic lineage. For each application/category, destination-service, or rule grouping, 20% of whole groups form the test set; the retained rows are divided into 60% training and 20% calibration, with the calibration portion held aside so that LightGBM, XGBoost, ordinal CatBoost, and native-categorical CatBoost are compared on identical training and test rows. The OOV rate is the share of test rows for which at least one grouping-field category is absent from training. This distinguishes unseen combinations from the stricter case in which a constituent category is itself unseen. The four lineages and the comparisons each supports are recorded in Table 4.

Uncertainty diagnostics use predictive entropy, multiclass Brier score, and 10-bin ECE. Let C be the target-class set and $\hat{p}(x) = (\hat{p}_c(x))_{c \in C}$ the predicted class-probability vector. Predictive entropy is

$$H(x) = -\sum_{c \in C} \hat{p}_c(x) \log_2 \hat{p}_c(x), \tag{5}$$

with $0 \log_2 0 = 0$ and normalized entropy obtained by dividing by $\log_2 |C|$ [50]. The Brier score is computed in the standard multiclass sum-over-classes form with range $[0, 2]$. ECE partitions the n test records into $B = 10$ equal-width bins $\{\mathcal{B}_b\}_{b=1}^{B}$ on confidence $\hat{c}(x_i) = \max_{c \in C} \hat{p}_c(x_i)$ and reports the

sample-size-weighted confidence-accuracy gap:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{B}_b|}{n} \Big| \operatorname{acc}(\mathcal{B}_b) - \operatorname{conf}(\mathcal{B}_b) \Big|, \tag{6}$$

where $\operatorname{acc}(\mathcal{B}_b)$ and $\operatorname{conf}(\mathcal{B}_b)$ are the bin accuracy and mean confidence. Because the target is near-deterministic, the vast majority of predictions concentrate in the top bin. As a result, 10 equal-width bins may leave lower-confidence bins sparsely populated, meaning the reported ECE effectively measures the top-bin confidence-accuracy gap and not uniform miscalibration.

Adaptive Prediction Sets (APS) complement point predictions with set-valued outputs that enjoy finite-sample, distribution-free marginal coverage under exchangeability of calibration and test data [46, 48]; this guarantee is not invoked for the temporal and context-held-out diagnostics, as explained below. Let $\mathcal{I}_{\mathrm{cal}}$ index the n_{cal} calibration records. Using the calibration set, the non-conformity score E_i for each instance i is computed by summing the predicted probabilities of all classes ranked higher than or equal to the true class y_i:

$$E_i = \sum_{c \in C : \hat{p}_c(x_i) \geq \hat{p}_{y_i}(x_i)} \hat{p}_c(x_i). \tag{7}$$

For a target error rate $\alpha \in (0,1)$, define $k_\alpha = \min\{n_{\mathrm{cal}}, \lceil (n_{\mathrm{cal}}+1)(1-\alpha) \rceil\}$ and let $\hat{q}_\alpha$ be the k_α-th smallest calibration score in $\{E_i\}_{i \in \mathcal{I}_{\mathrm{cal}}}$ from Eq. (7). For a test input x, let $c_{(1)}, \ldots, c_{(|C|)}$ order the classes by decreasing predicted probability. The implemented APS diagnostic retains classes whose cumulative mass up to that class is no larger than $\hat{q}_\alpha$ and falls back to the top-ranked class if this rule would return an empty operational set:

$$\tilde{C}_\alpha(x) = \Big\{ c \in C : \sum_{c' \in C:\, \hat{p}_{c'}(x) \geq \hat{p}_c(x)} \hat{p}_{c'}(x) \;\leq\; \hat{q}_\alpha \Big\}, \qquad \hat{C}_\alpha^{\mathrm{APS}}(x) = \begin{cases} \tilde{C}_\alpha(x), & \tilde{C}_\alpha(x) \neq \emptyset, \\ \{c_{(1)}\}, & \tilde{C}_\alpha(x) = \emptyset. \end{cases} \tag{8}$$

Probability ties are handled as a block in both calibration-score and candidate-set construction; the fixed encoded class order is used only to resolve a top-probability tie when the nonempty fallback is invoked. Under exchangeability of the calibration and test scores, the set of Eq. (8) satisfies the finite-sample distribution-free marginal coverage guarantee [46, 48]

$$\Pr\big(Y \in \hat{C}_\alpha(X)\big) \;\geq\; 1 - \alpha. \tag{9}$$

The analysis uses the deterministic, nonempty variant of the APS score. Without the randomization term of [48], and with the top-class fallback in Eq. (8), its sets can be conservative. APS set size is therefore treated as a model- and variant-sensitive expansion diagnostic, not as a pure measure of data ambiguity or a sharper theoretical claim than Eq. (9).

Completing the diagnostic suite, selective classification couples a classifier with a confidence-based selection (acceptance) rule $g_\lambda(x) = \mathbf{1}\{\hat{c}(x) \geq \lambda\}$, where $\hat{c}(x) = \max_{c \in C} \hat{p}_c(x)$ is the confidence of Eq. (6) and λ is the confidence threshold; the rule abstains when $g_\lambda(x) = 0$. Following the reject-option tradition in classification [51], the empirical retained coverage and selective risk on the n test records are

$$\widehat{\mathrm{cov}}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{c}(x_i) \geq \lambda\}, \qquad \widehat{\mathcal{R}}(\lambda) = \frac{\sum_{i=1}^{n} \mathbf{1}\{\hat{c}(x_i) \geq \lambda\}\, \mathbf{1}\{\hat{y}_i \neq y_i\}}{\sum_{i=1}^{n} \mathbf{1}\{\hat{c}(x_i) \geq \lambda\}}, \tag{10}$$

so that $\widehat{\mathcal{R}}(\lambda)$ is the error rate over the retained subset and $1 - \widehat{\mathrm{cov}}(\lambda)$ is the abstention rate. The selective risk is evaluated only when at least one record is retained. These formulations separate deterministic proxy reconstruction from regions of high operational ambiguity.

Because several scores are at or near the numerical ceiling, the Results section also reports error counts and confusion-matrix patterns. These counts are more informative than rounded accuracy alone. No synthetic oversampling is used in the main experiments, because the aim is to measure how reconstructable the logged operational decision is under the observed distribution instead of optimizing minority-class learning through generated samples.

For the main hold-out setting, the reported confusion matrix uses the label order Allow, Deny, and Drop. This order is kept consistent with the JSON outputs in the reproducibility package.

4.2 Uncertainty, Conformal, and Selective Diagnostics

Throughout the diagnostics, residual uncertainty denotes graded probabilistic indicators (confidence, entropy, calibration), residual ambiguity denotes non-singleton or empty set-valued reconstructions, and error pockets denote concentrated high-confidence error groups. LightGBM supplies the common feature-view lineage because it matches the full-data ceiling and can be run consistently across settings. The full-data benchmark fits support Tables 5, 8, and 9; the 80/20 diagnostic-pass fits support Tables 12 and 14; and the original conformal/selective lineage retrains LightGBM on 629,145 records. Small error-total differences between these lineages are reported where they affect interpretation.

Conformal sets are used as set-valued reconstructions of operational decisions, as opposed to causal security explanations. Because both conformal constructions yield nested sets as α decreases, their outputs can be read diagnostically as level sets of a consonant plausibility ordering over the three decisions [49]; entropy and confidence are likewise task-specific uncertainty summaries rather than universal measures of ambiguity [50]. A separate calibration split was used in the random and temporal train/calibration/test settings (629,145 training, 209,715 calibration, 209,716 test records); the 838,860-record training partition of the main 80/20 hold-out is further divided into these 629,145 training and 209,715 calibration records, while the same 209,716 test records are retained. In the context-held-out settings, calibration records are drawn only from retained contexts, so calibration and test populations differ by construction.

A paired application/category lineage compares ordinal LightGBM with native-categorical CatBoost on identical group-held-out rows. Retained contexts are divided into 638,486 training, 106,414 temperature-fitting, and 106,415 conformal-calibration records, followed by 197,261 test records with no application/category group overlap. Scalar temperature is fitted only on the first calibration subset; probability-threshold and deterministic APS quantiles use the second. Raw and temperature-scaled NLL, Brier score, fixed-bin ECE, adaptive ECE, conformal sets, and selective queues are reported without reusing calibration labels.

The main conformal score is the standard probability-threshold nonconformity score, one minus the predicted probability assigned to the true class [47], evaluated on each calibration record $i \in \mathcal{I}_{\text{cal}}$:

$$s_i = 1 - \hat{p}_{y_i}(x_i). \tag{11}$$

Using the same exact order-statistic index k_α defined above, let $\hat{q}$ be the k_α-th smallest calibration score in $\{s_i\}_{i \in \mathcal{I}_{\text{cal}}}$ of Eq. (11). The probability-threshold prediction set retains every class whose predicted probability is at least $1 - \hat{q}$:

$$\hat{C}_\alpha(x) = \left\{ c \in C : \hat{p}_c(x) \geq 1 - \hat{q} \right\}, \tag{12}$$

which satisfies the same marginal coverage guarantee as Eq. (9) under exchangeability [46, 47]. The resulting set of Eq. (12) can be read as a singleton-or-abstain reconstruction. Singleton sets mark near-deterministic regions, while empty sets mark cases where no class passes the calibrated probability threshold. At the $\alpha = 0.05$ operating point shown in Figure 3, the largest calibrated $\hat{q}$ among the displayed probability-threshold settings is 0.3772, giving a minimum threshold of $1 - \hat{q} = 0.6228$. Consequently, at most one class can pass in these displayed settings, and their probability-threshold sets are singleton or empty. At $\alpha = 0.01$, lower thresholds can admit multi-class sets in some feature-restricted settings; the full α sweep is therefore interpreted separately.

Marginal coverage is averaged over all test records and can hide undercoverage of rare classes. A Mondrian, class-conditional variant therefore calibrates a separate quantile $\hat{q}^{(c)}$ on the calibration scores restricted to each true class c, $\{s_i : i \in \mathcal{I}_{\text{cal}}, y_i = c\}$, and applies the class-specific threshold $1 - \hat{q}^{(c)}$ in Eq. (12), restoring the coverage guarantee $\Pr(Y \in \hat{C}_\alpha(X) \mid Y = c) \geq 1 - \alpha$ within every class under within-class exchangeability [46, 47]. Both variants are reported in the Results section.

The level $\alpha = 0.05$ is a conventional operating point. The Results section also reports how coverage moves across $\alpha \in \{0.01, 0.05, 0.10\}$, and the diagnostic conclusions rest on comparisons across settings at a common level, not on the specific level. An APS-style cumulative conformal set was also computed to show set expansion under ambiguity. The APS output is used for interpretation of ambiguous multi-class regions, not as a new classifier.

Selective classification was used to quantify operational abstention [51]. For confidence thresholds from 0.50 to 0.9999, the analysis reports retained coverage, abstention rate, retained accuracy, selective risk, retained errors, rejected errors, error-capture rate, and class-wise retained coverage. This allows the paper to ask whether high-confidence decisions remain reliable and whether low-confidence rejection captures error pockets.

The context-held-out conformal results require a careful interpretation. Random train/calibration/test conformal prediction has the cleanest exchangeability-based meaning, although even here exchangeability is only approximate, because all records are drawn from a single sub-hour window in which sessions are temporally correlated, so the random-split coverage guarantee should be read as near-nominal rather than exact. Context-held-out and temporal settings intentionally stress that assumption further. Their conformal outputs are therefore reported as diagnostic boundary evidence for probability and selective behavior under operational shift, without invoking formal finite-sample marginal coverage guarantees.

4.3 Experimental Environment and Reproducibility

All timing-sensitive experiments were run on a cloud-hosted aarch64 Linux virtual machine with four allocated CPU cores and 24 GB RAM using Python 3.12.3. The models were trained with xgboost 3.2.0, lightgbm 4.6.0, scikit-learn 1.8.0, numpy 2.4.4, pandas 3.0.3, scipy 1.17.1, and catboost 1.2.10; the complete pinned environment accompanies the released code in the repository requirements file so that the reported error counts can be reproduced. The same hardware environment was used for the final full-data benchmark, sample benchmark, ablation, and cross-validation results reported in this paper.

Wall-clock timings for the full queue and the individual benchmarks are reported in Section 5.9, and the original benchmark values in the Results section are taken from the released aggregate outputs. The review-motivated support, OOV, model/encoding, and shallow-tree checks form a distinct aggregate-only lineage run under Windows 11 with Python 3.12.12, xgboost 3.1.3, lightgbm 4.6.0, scikit-learn 1.8.0, numpy 2.3.5, pandas 2.2.3, and catboost 1.2.8. Their purpose is sensitivity analysis rather than replacement of the cloud timing results.

The determinism census, the Mondrian conformal comparison, and the residual-pocket breakdown are reproduced by dedicated scripts in the public repository, each emitting aggregate artifacts only. Because the restricted data cannot be redistributed, the repository also ships a runnable census-mechanics demonstration on a fully synthetic export built from an explicit, published, fictional policy rule; it lets a reader execute the determinism, conditional-entropy, and exhaustive determining-key census code end to end and confirm that it recovers the expected minimum determining keys where determinism holds, without exposing or reproducing the enterprise export's specific proxy structure. Unless otherwise stated, all stochastic steps use random seed 42, including the stratified 80/20 hold-out, the capped-sample per-class subsampling, the five-fold stratified cross-validation, the context-held-out group partitioning that holds out whole context groups, and the label-permutation reference, so that the reported counts regenerate deterministically.

5 Results

5.1 Core Hold-Out Reconstruction

The full-data core baseline addresses the initial reconstruction aspect of *RQ1*, demonstrating that near-perfect accuracy is attainable even when direct label-source fields are excluded. As detailed in Table 5, both XGBoost and LightGBM reached perfect hold-out performance. Extra Trees made one error in 209,716 test records, and CatBoost remained very close to perfect. Cross-validation, group-disjoint, and multi-seed summaries below assess within-export split sensitivity; none establishes row independence or temporal generalization.

Table 5 Full-data core hold-out benchmark.

Model	Accuracy	Bal. acc.	Macro-F1	Errors	Fit (s)	Pred. (s)	Pred. ms/1k
XGBoost	1.000000	1.000000	1.000000	0	25.613	1.233	5.878
LightGBM	1.000000	1.000000	1.000000	0	24.500	4.029	19.210
Extra Trees	0.999995	0.999998	0.999939	1	45.324	1.127	5.374
CatBoost	0.999905	0.999959	0.998946	20	45.722	0.754	3.595

Here the central point is interpretive, not competitive. Multiple models reconstruct the target with near-perfect fidelity, and two achieve exact reconstruction. Because direct label-source fields were excluded and preprocessing was confined to the training pipeline, the outcome points to redundant encoding across the remaining structured context and not simple direct-label leakage. Part of this core

signal nonetheless comes from volume and duration fields that are recorded at session end and are therefore partly post-decisional (Section 3.4); this perfect core benchmark is accordingly best read as a proxy-rich operational upper bound. The more conservative without-volume/duration view leaves only 38 conflicting records and reconstructs at 0.9962 macro-F1, but it remains a reduced telemetry view rather than a verified decision-time feature set.

A duplicate-structure check reinforces this interpretation. The 1,048,576 records collapse to 838,202 distinct core feature vectors, and no distinct vector carries more than one decision label. Across 20 prespecified label permutations with the same class marginals, conflicted contexts range from 27,413 to 27,927 (mean 27,694), and conflict rows range from 193,043 to 194,342 (mean 193,811). This is a descriptive null check, not an inferential permutation test. Repetition alone also fails to explain the scores, since only 24.31% of test records have an exact feature-vector match in the training partition, and exact-match lookup with majority fallback reaches 0.868 accuracy and 0.590 macro-F1. The learned reconstruction therefore generalizes across distinct context combinations within the export. Uncertainty under feature restriction or context shift is evidence-relative and consistent with an epistemic interpretation, but the export does not determine the uncertainty type outside its observed contexts [52].

The fixed-depth tree sensitivity provides a deliberately shallow counterpoint to the ensemble ceiling. Across the complete prespecified curve, depth 4 reaches 0.9409 macro-F1 with 1,362 errors and 13 leaves; depths 6, 8, and 12 reach 0.9906, 0.9959, and 0.9967 macro-F1 with 158, 68, and 55 errors, respectively. No depth is selected on the test result. Compact trees therefore approximate most of the decision surface, while the last fraction requires richer partitioning.

5.2 Determinism Profile and Information Limits Across Feature Views

The model-free audit in Table 6 covers the eight random feature views, pairing the conditional entropy of Eq. (2) and the in-sample resubstitution error of Eq. (3) with the corresponding LightGBM test errors from the uncertainty diagnostics.

Determinism does not depend on application metadata. The 19-field no-application view still carries zero conflict rows (equivalently $H(Y \mid V) = 0$ by Eq. (4)), so the decision label remains an exact function of the remaining context. Even removing all volume and duration fields leaves only 38 conflicting records among 1,048,576.

The restricted views clarify where information is lost. The zone/interface-only view collapses to 24 distinct contexts whose in-sample resubstitution error is exactly 13,809 records, the size of the Deny class. Interface placement separates Allow from Drop perfectly but cannot see Deny at all.

Comparing lookup against learning then locates the contribution of model generalization. The lookup witness attains the scaled floor exactly in the zone/interface view (2,762 test errors against the scaled 13,809-record limit) and closely tracks it in the other low-cardinality views, whereas the balanced-weighted LightGBM deliberately trades majority accuracy for minority recall and lands far above it there (30,514 errors). In the rich views, the relation reverses, with the lookup stranded at 27,590 test errors because only 24.31% of test records have an exact training match, while the ensembles generalize across the 75.7% of test records with unseen contexts at near-zero loss.

Table 6 Feature-view determinism and information audit.

View	Fields	Contexts	Conflicts	$H(Y \mid V)$	In-sample err.	Lookup err.	Model err.
Core all features	24	838,202	0	0.000000	0	27,590	2
Without application context	19	838,195	0	0.000000	0	27,589	7
Without volume/duration	17	504,992	38	0.000034	16	26,946	62
Transport + volume only	10	830,322	17,711	0.004659	1,406	26,394	1,168
Only application context	5	342	346,717	0.042950	9,472	1,904	2,381
Only zone/interface	4	24	210,002	0.064523	13,809	2,762	30,514
Only transport/ports	3	343,554	188,287	0.104918	32,802	30,208	19,635
Only volume/duration	7	162,339	232,648	0.046926	11,870	2,411	3,424

The full-export census finds no determining single field, pair, or triple among 2,324 candidates and exactly seven determining four-field sets among 10,626 candidates. Every set contains Source Port and session-end Bytes, completed by two interface, application-taxonomy, country, or destination-port fields. The exact sets are close to unique-column combinations: $|\mathcal{V}|/N$ is 0.721–0.779, 89.74–91.59% of distinct contexts are singletons, and 64.70–71.35% of records occupy singleton contexts. The repeated-context

mass remains material, however; records in contexts with support of at least 2, 5, and 10 account for 28.65–35.30%, 14.84–20.51%, and 11.96–13.56% of the export, with zero observed conflicts in every stratum (Table 7).

Across 20 cardinality-preserving Source Port permutations per key, none of the 140 controls remains exact; the controls produce 7–59 majority-label errors. This rules out automatic exactness from cardinality alone, but the small absolute gap does not make Source Port a causal policy variable. Source Port alone incurs 177,286 majority-label errors, and 94.32% of records use ports observed under at least two labels, so it behaves as a shared session fingerprint rather than a stand-alone rule. Using the IANA System, User, and Dynamic/Private ranges [62], dynamic/private ports account for 68.18% of Allow, 70.22% of Deny, and 52.11% of Drop records. The value 0 occurs in 4.87% of records (5.37% of Allow, 0% of Deny, and 3.40% of Drop); these descriptive differences do not establish a causal policy role.

Table 7 Support and train-to-test lookup coverage for the seven full-export determining keys.

Key	Fields beyond Source Port + Bytes	Singleton rows	$n \geq 2$	$n \geq 5$	$n \geq 10$	Strat. seen	Chron. seen
K1	Application; Outbound Interface	64.72%	35.28%	20.51%	13.56%	33.57%	29.95%
K2	Application; Destination Country	65.01%	34.99%	20.39%	13.52%	33.31%	29.72%
K3	Destination Port; Outbound Interface	71.35%	28.65%	14.84%	11.96%	26.88%	19.97%
K4	Outbound Interface; Subcategory of app	64.71%	35.29%	20.51%	13.56%	33.58%	29.95%
K5	Category of app; Outbound Interface	64.71%	35.29%	20.51%	13.56%	33.58%	29.95%
K6	Outbound Interface; Technology of app	64.70%	35.30%	20.51%	13.56%	33.59%	29.98%
K7	Outbound Interface; Risk of app	64.72%	35.28%	20.51%	13.56%	33.57%	29.95%

Notes. Singleton and support columns are shares of all records. Seen-test columns use a mapping fitted in the corresponding training block; all seen-context labels agree for K1–K7.

The training-only exhaustive census finds eight zero-conflict four-field candidates. Seven are K1–K7 and retain 100% agreement on seen contexts in both test protocols. The eighth candidate, which adds Source Country to Outbound Interface, Source Port, and Bytes, makes one seen-context mapping error in each protocol and therefore fails full-export determination. Discovery and stratified test validation are thus non-circular, while the limited 26.88–33.59% stratified and 19.97–29.98% chronological coverage prevents a stand-alone lookup-policy claim. In the 17-field view without volume and duration, the best one- through four-field subsets incur 13,599, 6,166, 1,567, and 301 empirical majority-label errors, respectively. The best four-field subset (Application, Outbound Interface, Source Port, and Category) reaches 99.9713% in-sample agreement but does not restore exact determination. Because every exact key includes the post-decision byte field, K1–K7 remain all-field export fingerprints rather than causal rule reconstructions.

Under restriction, the trained models approach but do not reach the view floors (for the transport-plus-volume view, an in-sample floor rate of 0.13% against a 0.56% model test error), so near-perfect core scores reflect the information structure of the export, and the specific contribution of gradient boosting is generalization across unseen contexts and not the determinism itself.

5.3 Full-Data With-Policy Setting

Reintroducing policy-related fields establishes an empirical upper boundary, though this addition did not materially alter the outcome. As shown in Table 8, XGBoost and LightGBM once again achieved perfect accuracy, balanced accuracy, and macro-F1.

Table 8 Full-data with-policy upper-bound benchmark.

Model	Accuracy	Bal. acc.	Macro-F1	Errors	Fit (s)	Pred. (s)	Pred. ms/1k
XGBoost	1.000000	1.000000	1.000000	0	25.092	1.239	5.908
LightGBM	1.000000	1.000000	1.000000	0	24.652	4.117	19.629
Extra Trees	0.999990	0.999996	0.999877	2	50.899	1.192	5.682
CatBoost	0.999962	0.999983	0.999623	8	46.385	0.848	4.046

The with-policy setting is an empirical upper bound, not the primary claim, since policy-related fields are intrinsically proximate to the decision logic. Because the core setting already operates at the

ceiling, the with-policy setting provides no headroom for additional signal; its role is to document that the ceiling persists when explicit policy context is included.

5.4 No-Application-Context Ablation

Addressing the proxy-dependence aspect of *RQ1*, the strongest cautionary result is the no-application-context ablation. After removing the application and category variables, XGBoost still achieved 0.999981 accuracy, 0.999987 balanced accuracy, and 0.999981 macro-F1 (Table 9), so suppressing application and category context reduces the ceiling only marginally.

Table 9 Full-data no-application-context ablation benchmark.

Model	Accuracy	Bal. acc.	Macro-F1	Errors	Fit (s)	Pred. (s)	Pred. ms/1k
XGBoost	0.999981	0.999987	0.999981	4	23.687	1.134	5.406
Extra Trees	0.999976	0.999984	0.999920	5	46.043	0.985	4.698
LightGBM	0.999967	0.999970	0.999911	7	22.505	4.208	20.066
CatBoost	0.999905	0.999959	0.999002	20	44.303	0.584	2.783

The no-application-context result is especially important for the minority Deny class because aggregate scores near one can still hide class-specific failures. The XGBoost confusion pattern under this ablation, in the benchmark output label order of Allow, Deny, and Drop, contains only four errors. Three Allow records were predicted as Drop, and one Drop record was predicted as Allow, while Deny precision, recall, and F1 are all 1.000. Reconstruction is therefore not anchored only to explicit application signatures; transport, volume, and zone context redundantly encode enforcement outcomes, partly through the post-decision fields identified in Section 3.4. Because transport and volumetric proxies remain, this is not a proxy-free test. These fields are valid operational measurements, but they are also traces of the system that produced the label.

5.5 Capped-Sample Cross-Validation

Cross-validation on the capped sample checks split stability within the same operational export. As shown in Table 10, all four top models achieved mean macro-F1 above 0.9997, reported with across-fold standard deviations, not a single threshold, so the stability is relative rather than absolute. LightGBM attained the highest mean macro-F1, followed closely by Extra Trees, XGBoost, and CatBoost. The across-fold standard deviations were small, indicating that this performance is grounded in the data structure and not a single lucky split.

Table 10 Capped-sample five-fold cross-validation stability.

Model	Mean accuracy	Mean balanced accuracy	Mean macro-F1	Mean weighted F1
LightGBM	0.999947 ± 0.000043	0.999908 ± 0.000100	0.999908 ± 0.000075	0.999947 ± 0.000043
Extra Trees	0.999930 ± 0.000060	0.999824 ± 0.000162	0.999877 ± 0.000104	0.999930 ± 0.000060
XGBoost	0.999877 ± 0.000070	0.999784 ± 0.000053	0.999811 ± 0.000071	0.999877 ± 0.000070
CatBoost	0.999859 ± 0.000090	0.999823 ± 0.000122	0.999780 ± 0.000110	0.999859 ± 0.000090

A full-data multi-seed check in the reproducibility package agrees with this pattern. Re-running the LightGBM core hold-out across ten alternative split seeds on the experiment environment yields a mean macro-F1 of 0.99985 ± 0.00011 (minimum 0.99969, maximum 1.000000, and a mean of 2.4 errors in 209,716 test records), so the near-perfect core result is not specific to the benchmark seed. Because the leading models sit at or against the numerical ceiling in this setting, pairwise significance testing between them is uninformative, so the cross-validation and multi-seed evidence is used to establish split stability instead of ranking models. However, cross-validation within the same operational export cannot prove generalization to different organizations, policies, time windows, or firewall configurations. It answers the narrower question of whether the signal is stable within this export and sampling regime.

Another complementary session-group check aligned the raw Session ID field with the processed rows with zero action mismatches and repeated the core LightGBM hold-out over five 80/20 group-disjoint

partitions. Group overlap was zero in every run. Only 1,521 nonzero Session IDs repeated, spanning 3,067 records with at most three records per nonzero session; zero-valued sentinel IDs were assigned row-unique surrogate groups. Macro-F1 averaged 0.999775 (SD 0.000033; range 0.999749–0.999812), with three to four errors per split. Same-Session-ID overlap therefore does not explain the near-ceiling result in this export, although this check does not establish independence across organizations, policy periods, or longer time horizons.

5.6 External Instantiation of Schema-Independent Components

The within-export checks above cannot show whether the reconstruction procedure itself transfers beyond this deployment. To probe this, the schema-independent protocol components are instantiated on an independent public firewall dataset with a different schema. The UCI Internet Firewall dataset [63], also studied by [1, 14], comprises 65,532 sessions labeled Allow, Deny, Drop, or Reset-both and is described only by transport and volume fields (source, destination, and NAT ports; bytes and packets sent and received; and elapsed time). This dataset exposes no rule identifier, timestamps, zone or application signatures, or threat descriptors, so the policy-conditioned components of the protocol (the field-level determinism audit over a rich context, the descriptor and context-held-out ablations, and the chronological stress test) are not applicable and are not claimed to transfer. The evaluated components are core hold-out reconstruction, exact-duplicate structure, a duplicate-aware grouped split, and cross-validation stability, all run with the same model configurations and preprocessing pipeline as the primary experiments; their reconstruction scores against the majority-class baseline are reported in Table 11.

Table 11 Public UCI instantiation of schema-independent reconstruction components.

Model / probe	Accuracy	Macro-F1
Decision Tree	0.9979	0.8673
Random Forest	0.9978	0.8550
Extra Trees	0.9977	0.8479
XGBoost	0.9982	0.8422
LightGBM	0.9981	0.8913
CatBoost	0.9963	0.8402
XGBoost duplicate-aware grouped split (median of 5)	—	0.8643
Majority-class baseline	—	0.1824

Three patterns from the enterprise export reappear. Accuracy is again near-perfect (0.996–0.998), and exact-duplicate signatures account for 20.3% of records, so the redundant, near-deterministic structure is not unique to the Palo Alto export. Macro-F1 is markedly lower (0.84–0.89) because the Reset-both class has only 54 records; for XGBoost its per-class F1 is 0.375 (recall 0.273 on eleven test cases) while Allow, Deny, and Drop stay at 0.996–1.000, so the same accuracy-versus-minority gap that motivates the class-conditional treatment in Section 5.8 recurs on an independent dataset. The duplicate-aware grouped split, which places identical signatures entirely on one side, still yields a median macro-F1 of 0.864 (five-fold cross-validation macro-F1 0.90 for LightGBM and 0.85 for XGBoost), far above the 0.182 majority-class baseline, so reconstruction is not an artifact of duplicate memorization. This second instantiation corroborates the evaluated schema-independent patterns without establishing transfer of the enterprise policy-conditioned layer, whose proxy structure remains specific to the primary export.

5.7 Temporal and Feature-Family Diagnostics

The LightGBM diagnostic pass tests whether the near-deterministic result depends on random splitting or on one narrow feature family; these are targeted stress tests, not replacements for the core benchmark. The core recheck has two residual errors, versus zero in Table 5, which is within the seed-level variation of Section 5.5. The behavior of macro-F1 and balanced accuracy when temporal bounds or entire feature families are removed is reported in Table 12.

Under the time-ordered split, performance remains near the ceiling despite using later records for hold-out; given the single sub-hour span, this is a short-term, within-window stability check against high temporal autocorrelation, as opposed to a longitudinal concept-drift validation. Because the export contains no diurnal or multi-day variation, genuine concept drift cannot be assessed here at all, and

Table 12 LightGBM temporal and feature-family diagnostics.

Diagnostic setting	Split	Features	Accuracy	Bal. acc.	Macro-F1	Errors
Core recheck	Random 80/20	24	0.999990	0.999996	0.999877	2
Core within-window time-ordered	Time-ord. 80/20	24	0.999971	0.999729	0.999599	6
Without application context	Random 80/20	19	0.999967	0.999738	0.999797	7
Without zone/interface fields	Random 80/20	20	0.999981	0.999873	0.999755	4
Without transport/port fields	Random 80/20	21	0.999995	0.999998	0.999939	1
Without country fields	Random 80/20	22	0.999990	0.999996	0.999877	2
Without volume/duration fields	Random 80/20	17	0.999704	0.998567	0.996221	62
Without risk/SaaS/AI fields	Random 80/20	21	0.999990	0.999996	0.999877	2
Transport plus volume only	Random 80/20	10	0.994431	0.994603	0.983327	1,168

the time-ordered result should be read only as an autocorrelation-robustness check, not as evidence of temporal generalization. With that scope, it is still stronger evidence than a purely random split that the learned signal persists across the export's short time interval.

The feature-family results in Table 12 reinforce the proxy-rich interpretation. Removing application, zone/interface, country, risk/SaaS/AI, or transport/port fields barely changes the outcome, while removing volume and duration gives the largest leave-one-family-out drop yet stays high. A five-seed rerun confirms that this is not a one-split artifact: the core LightGBM mean macro-F1 is 0.999828, and the without-volume/duration mean is 0.996186, with all other single-family removals remaining above 0.9996. Following Section 3.4, the without-volume/duration setting is treated as a reduced telemetry view, not a verified decision-time feature set. Normalized-gain diagnostics also rank volume/duration and zone/interface highly, but the corresponding figure is omitted because high-cardinality split importances are representation-sensitive and do not support causal interpretation.

To separate ordinary leave-one-family-out ablation from a more severe proxy-minimal stress test, an additional full-data check retained only zone/interface and transport/port fields, optionally adding country context. This view removes application/category, volume/duration, policy/rule/action-source, risk/SaaS/AI, raw label-source, and timestamp fields. It is deliberately conservative and should not be read as a causal deconfounding design; its role is to test whether reconstructability collapses when several prominent proxy families are removed at once.

Table 13 Strict proxy-minimal stress test on full-data 80/20 hold-outs.

Feature view	Model	Macro-F1 mean (SD)	Macro-F1 range	Error range
Zone/interface + transport/ports	XGBoost	0.951301 (0.001283)	0.949480–0.952833	792–862
+ country context	XGBoost	0.962710 (0.001253)	0.961306–0.964388	591–652
Zone/interface + transport/ports	LightGBM	0.804635 (0.212445)	0.424643–0.903022	2,172–95,577
+ country context	LightGBM	0.859041 (0.136098)	0.615639–0.924486	1,559–34,905

The strict-view results in Table 13 sharpen the boundary. The strict view no longer sits at the numerical ceiling, which is useful evidence rather than a weakness; XGBoost remains stable at approximately 0.951 macro-F1 without country context and approximately 0.963 with country context, whereas LightGBM becomes seed-sensitive in the sparse view. The defensible reading is therefore not a model leaderboard but a boundary statement. Even severe proxy removal leaves substantial reconstructability, while sparse-view instability prevents any claim of proxy-free determinism.

5.8 Uncertainty, Conformal, and Selective Diagnostic Results

The uncertainty analysis asks whether near-deterministic reconstruction also implies low uncertainty (*RQ3*) and completes the context-held-out side of *RQ2*. Expected calibration error, multiclass Brier score, and mean predictive entropy for representative feature views and stress tests are reported in Table 14. The core random setting has only two errors in 209,716 test records, with ECE 0.000010, Brier score 0.000019, and mean entropy 0.000004 bits; the temporal core setting stays similarly low at six errors, ECE 0.000029, Brier 0.000056, and mean entropy 0.000008 bits.

Context-held-out estimates are single-partition diagnostics whose magnitude reflects unfamiliar context, OOV handling, model objective, and representation rather than a partition-free data property. The

Table 14 Uncertainty and calibration diagnostics by feature view.

Diagnostic setting	Split	Macro-F1	Errors	ECE	Brier	Mean entropy
Core all features	Random	0.999877	2	0.000010	0.000019	0.000004
Without volume/duration	Random	0.996221	62	0.000393	0.000657	0.001304
Transport + volume only	Random	0.983327	1,168	0.002285	0.008586	0.016122
Only application context	Random	0.941829	2,381	0.061662	0.047918	0.230375
Only zone/interface	Random	0.682252	30,514	0.116886	0.193182	0.122222
Temporal core all features	Temporal	0.999599	6	0.000029	0.000056	0.000008
Temporal transport + volume only	Temporal	0.978377	1,152	0.002140	0.008407	0.015471
App/category held out	Group-held-out	0.673224	5,658	0.028179	0.054106	0.013998
Dest. service held out	Group-held-out	0.805951	3,513	0.005945	0.012745	0.005043
Rule held out	Group-held-out	0.927519	3,446	0.005021	0.010294	0.008645

Table 15 Paired seed-42 context-held-out macro-F1 across model and encoding choices.

Held-out context	Row OOV	LGB ord.	XGB ord.	Cat ord.	Cat native
Application/category	0.9575	0.6725	0.6858	0.6828	0.9923
Destination service	0.5277	0.7991	0.8040	0.7985	0.7716
Rule	1.0000	0.9423	0.7121	0.8895	0.8794

Notes. All models use identical training and test rows. Row OOV is the test-row share with an unseen grouping-field category; the paired results diagnose representation sensitivity, not algorithm ranking.

application/category estimate differs slightly between diagnostic lineages (0.673224 with 5,658 errors versus 0.672548 with 5,664), within the partition variation reported below.

On identical application/category rows, the three ordinal pipelines cluster at 0.673–0.686 macro-F1, while native-categorical CatBoost reaches 0.9923. Recovery is not uniform because native CatBoost is lower for destination-service and rule hold-outs; unfamiliar context alone therefore cannot explain the observed boundary (Table 15).

Table 16 Paired application/category held-out diagnostics for ordinal LightGBM and native CatBoost.

Pipeline	Prob.	Macro-F1	F1 A/Dn/Dr	aECE	PT cov.	APS size	Queue	Captured / total
LGB ordinal	Raw	0.6725	.9849/.0328/1	.02765	.3156	1.7346	.1772	2,371 / 5,664
LGB ordinal	Temp.	0.6725	.9849/.0328/1	.02538	.3037	1.6953	.2388	5,663 / 5,664
Cat native	Raw	0.9923	.9996/.9774/1	.002634	.1280	1.9484	.1579	154 / 154
Cat native	Temp.	0.9923	.9996/.9774/1	.000144	.1285	1.9672	.0526	107 / 154

Notes. All rows use identical training and test records; scalar temperature fitting and conformal calibration use separate calibration subsets. PT coverage and APS size are empirical at $\alpha = 0.05$, and queue values use threshold 0.999. No coverage guarantee is invoked under context shift.

Scalar temperature fitting leaves point predictions unchanged and reduces test NLL/Brier/adaptive ECE from 0.1900/0.05614/0.02765 to 0.1265/0.05428/0.02538 for ordinal LightGBM and from 0.00542/0.001497/0.002634 to 0.00373/0.001312/0.000144 for native CatBoost. Under context shift, probability-threshold coverage at $\alpha = 0.05$ is 31.56% overall and 1.56% for Deny with raw ordinal LightGBM, versus 12.80% overall and 0.03% for Deny with raw native CatBoost; deterministic APS reaches 100% empirical coverage only with average set sizes of 1.735 and 1.948. A fixed 0.999 confidence threshold is not a model-invariant queue rule. Temperature scaling raises ordinal-LightGBM error capture from 41.86% to 99.98% while increasing the queue from 17.72% to 23.88%, but lowers native-CatBoost capture from 100% to 69.48% while reducing the queue from 15.79% to 5.26%.

Ordinal zone- and country-pair hold-outs remain near the ceiling (macro-F1 0.995850 and 0.991891; APS average set size 1.209 and 1.147), while destination-service hold-out remains difficult (0.799132; APS size 1.650). The precision, recall, and F1 values for the three ordinal-LightGBM context definitions are shown in Figure 2; Deny F1 is 0.0348 under application/category shift, 0.4215 under destination-service shift, and 0.7932 under rule shift, while Allow and Drop remain near the ceiling.

The point estimates in Table 16 and Figure 2 each rest on a single group partition. Five seeded LightGBM partitions show substantial variance (Table 17). Seed 42 reproduces the point estimates exactly but is not representative. Application/category macro-F1 ranges from 0.2532 to 0.9441 (mean 0.5561, SD 0.2654), and the zone-pair and country-pair hold-outs average 0.8873 and 0.9080 with minima of 0.6735 and 0.7140. These sampled means characterize the ordinal-LightGBM pipeline only; they do not

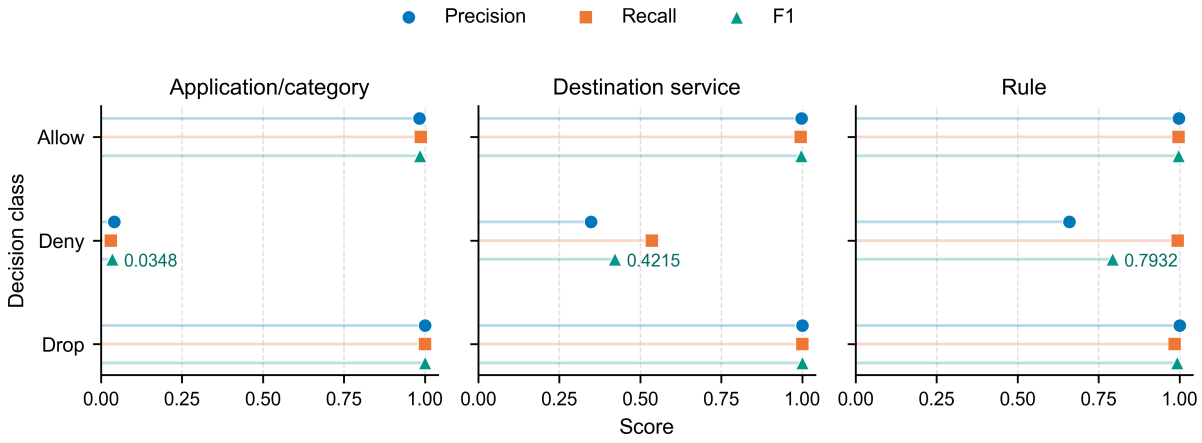


Fig. 2 Ordinal-LightGBM precision, recall, and F1 by class under application/category, destination-service, and rule context hold-outs

establish a partition-invariant or model-invariant ranking. Deny is least stable for application/category in that lineage, with mean F1 0.2899 (SD 0.3896).

Table 17 Five-seed partition variance of the LightGBM context-held-out diagnostics.

Held-out context	Single-part. macro-F1	Macro-F1 mean (SD)	Macro-F1 range	Deny F1 mean (SD)
Application/category	0.6725	0.5561 (0.2654)	0.2532–0.9441	0.2899 (0.3896)
Destination service	0.7991	0.7920 (0.1226)	0.7051–1.0000	0.3854 (0.3625)
Zone pair	0.9958	0.8873 (0.1355)	0.6735–0.9958	0.5067 (0.4779)
Country pair	0.9919	0.9080 (0.1139)	0.7140–0.9919	0.7253 (0.3418)

Excluding the 3,621 auth-policy-redirect records from the Deny class does not remove this brittleness. Under the same five-partition protocol, redirect-excluded application/category hold-out gives a Deny F1 mean of 0.3021 (versus 0.2899 with redirects) and macro-F1 mean 0.5637 (versus 0.5561), and destination-service hold-out gives Deny F1 mean 0.5415 (versus 0.3854); the minority collapse under context shift therefore reflects genuine boundary fragility rather than the labeling of redirected sessions, consistent with the earlier core-hold-out insensitivity to the same records.

Conformal prediction adds a set-valued perspective. At $\alpha = 0.05$, probability-threshold sets in the random core setting reach coverage 0.949651, singleton rate 0.949656, and empty-set rate 0.050344, so the output behaves as a singleton-or-abstain system. Under application/category held-out, the same rule gives coverage 0.315551 and empty rate 0.684449, with destination-service and rule held-out at intermediate coverage of 0.523947 and 0.516509. The calibrated in-distribution rule therefore remains near nominal for random and proxy-restricted views but does not transfer cleanly to unfamiliar operational contexts. At the displayed thresholds, this rule and confidence-based selective classification both act on the same top-class probabilities; they report different operating quantities but are not treated as independent corroborating evidence.

The α grid shows that the pattern is not a single-threshold artifact. Core coverage moves through 0.990, 0.950, and 0.902, while application/category held-out coverage moves through 0.722, 0.316, and 0.250. Marginal validity also hides a class-conditional deficit; within the core random setting, rare-Deny coverage under marginal probability-threshold sets falls to 0.932295, 0.834540, and 0.580377 for $\alpha = 0.01$, 0.05, and 0.10. The Mondrian variant, calibrated per class on the same 209,715-record split, restores Deny coverage to 0.996379, 0.951484, and 0.906951 (Table 18). This repair applies within the exchangeable regime; no Mondrian coverage claim is made for the context-held-out settings, where within-class exchangeability is deliberately broken.

Set expansion in the nonempty APS cumulative sets of Eq. (8) exposes ambiguity. At $\alpha = 0.05$, the random core setting has average set size 1.325693 and singleton rate 0.674307, versus 1.734798 and 0.265202 under ordinal-LightGBM application/category hold-out. A separate ordinal follow-up gives average set size 1.650 for destination-service shift, while zone- and country-pair shifts require much smaller expansions. Deny follows the same pattern, with average set size rising from 1.263577 in the random core setting to 1.983518 when application/category contexts are held out, and to 1.999494 in the rule-held-out diagnostic. These values characterize deterministic nonempty APS; Table 16 shows

Table 18 Marginal and Mondrian conformal coverage in the core random setting.

α	Calibration	Coverage	Empty rate	Allow	Deny	Drop
0.01	Marginal	0.989772	0.010223	0.988814	0.932295	0.996568
0.01	Mondrian	0.990521	0.009475	0.989535	0.996379	0.993614
0.05	Marginal	0.949651	0.050344	0.951267	0.834540	0.950908
0.05	Mondrian	0.949799	0.050196	0.949745	0.951484	0.949887
0.10	Marginal	0.901886	0.098114	0.904585	0.580377	0.911743
0.10	Mondrian	0.901967	0.098028	0.901422	0.906951	0.903575

that expansion also depends on model representation and probability scaling. The singleton-or-empty probability-threshold view is placed next to the nonempty APS view in Figure 3.

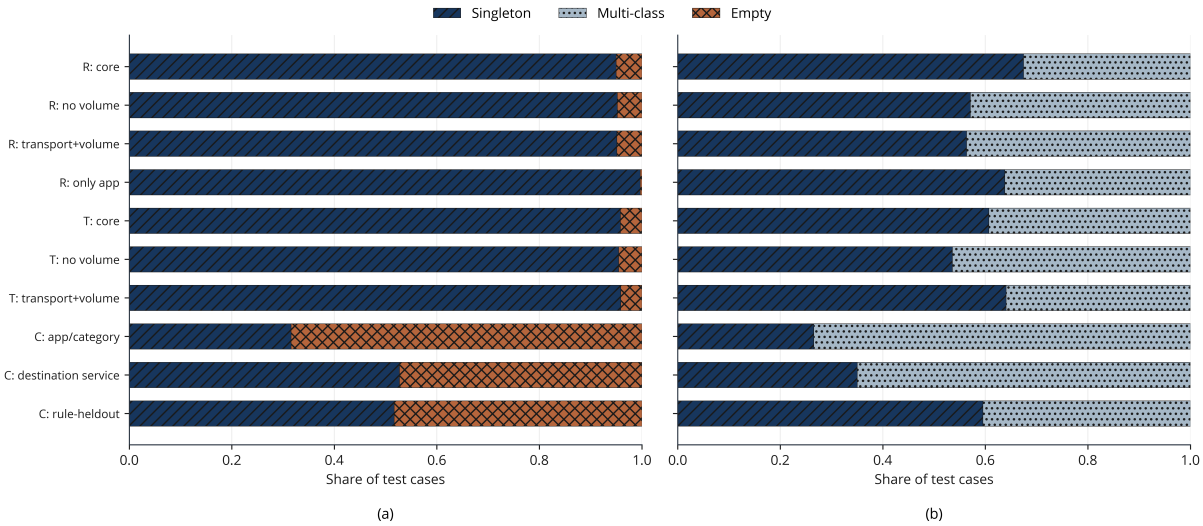


Fig. 3 Probability-threshold (a) and cumulative APS (b) shares of singleton, multi-class, and empty prediction sets across random, temporal, and context-held-out settings at $\alpha = 0.05$

Selective classification shows the retained-coverage and risk trade-off within a fixed probability lineage. At threshold 0.999, raw ordinal LightGBM retains 82.2788% of application/category test records and captures 41.8609% of its errors. Temperature scaling and native categorical modeling change both quantities materially (Table 16), so a fixed confidence threshold is not representation- or calibration-invariant. The raw ordinal summary below is a diagnostic comparator, not an autonomous queue rule.

Table 19 Raw ordinal-LightGBM confidence-only queues at threshold 0.999.

Diagnostic setting	Queue rate	Selective risk	Captured / total errors	Dominant pocket (80/20 fit)
Core all features	0.0062%	0.000005	2 / 3 (66.67%)	Allow → Deny (2 errors)
Transport + volume only	10.9467%	0.000198	1,148 / 1,185 (96.88%)	Allow → Drop (917 errors)
Temporal core all features	0.0024%	0.000024	2 / 7 (28.57%)	Allow → Deny (4 errors)
App/category held out	17.7212%	0.020289	2,371 / 5,664 (41.86%)	Deny → Allow (3,235 errors)
Destination service held out	7.1613%	0.004450	1,592 / 3,847 (41.38%)	Allow → Deny (2,403 errors)
Rule held out	2.5420%	0.002524	1,159 / 2,806 (41.30%)	Allow → Deny (2,025 errors)

The dominant residual pocket explains the raw ordinal application/category result. In the 80/20 fit, 3,235 Deny records are reconstructed as Allow with maximum predicted probability at least 0.99; 3,217 are ms-update sessions with auth-policy-redirect, so the failure is one policy interaction rather than diffuse noise. Context-held-out queue rates are batch-level diagnostics that call for representation and calibration checks, not steady-state per-record workloads. Core and temporal capture percentages in Table 19 use only a handful of total errors and are small-sample ratios. The raw ordinal retained-coverage and selective-risk trade-off is shown in Figure 4.

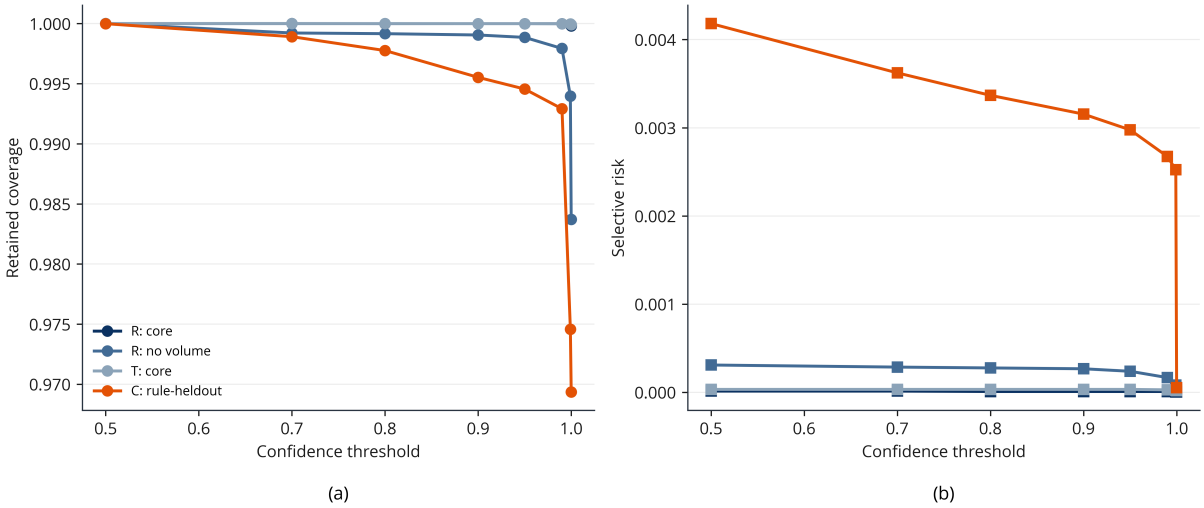


Fig. 4 Retained coverage (a) and selective risk (b) across confidence thresholds for random-core, proxy-restricted, temporal-core, and rule-held-out diagnostic settings

Calibration provides a complementary probability-quality view of the same boundary behavior. ECE stays near 10^{-5} for the core random and temporal settings (Table 14), whereas context-held-out settings carry larger high-confidence deviations. Because the near-deterministic target concentrates almost all probability mass in the top confidence bin, these near-10^{-5} core ECE values reflect the top-bin confidence-accuracy gap, not uniform calibration quality, so the calibration signal is most informative in the proxy-restricted and context-held-out views. Collectively, these diagnostics argue against reading the experiments as a model competition. The label is near-deterministic in the observed core feature space, but uncertainty appears under feature restriction, unfamiliar operational contexts, conformal abstention, APS set expansion, and selective-risk pockets.

5.9 Runtime and Scalability

Runtime results address the practical side of *RQ1*. In the evaluation environment described in Section 4.3, the full core benchmark completed in 155.94 seconds and the no-application-context ablation completed in 151.15 seconds. The complete experimental queue, including sample experiments and cross-validation, completed in 591 seconds.

Among the full-data settings, XGBoost gives the best observed balance of accuracy and runtime, training in roughly 24 to 26 seconds (Tables 5 and 9) with prediction latency near 5 to 6 ms per 1,000 rows; LightGBM trains slightly faster but predicts slower, and CatBoost predicts fastest at a lower macro-F1. These environment-specific values show that million-row decision reconstruction is computationally feasible on modest cloud hardware.

6 Discussion

The results are most informative when read as evidence about an operational label-generating process, not as a classifier leaderboard. XGBoost and LightGBM reconstruct the full core hold-out labels exactly, the no-application-context ablation stays near-perfect, and the model-free audit finds seven all-field determining sets that survive training-only discovery and seen-context test checks. A stricter proxy-minimal check no longer remains near-perfect. XGBoost is stable around 0.951 macro-F1 without country context and 0.963 with country context, whereas LightGBM is seed-sensitive. Substantial reconstructability therefore survives severe proxy removal for one established learner, but the magnitude is model-dependent. This pattern extends prior firewall-action studies that reported high learnability on firewall decision or action labels [2, 16, 64], but it changes the interpretation of such scores. The models show that the decision is attainable from the remaining structured context in this export; they do not establish a transferable attack detector, an independent security truth, or a new classification algorithm [5, 6].

Proxy interpretation is central to this reading. Direct label-source fields are excluded and preprocessing is confined to the training partition, which addresses the most obvious leakage pathways emphasized in reproducibility-oriented security-ML work [5, 40, 42]. Yet the remaining fields are still produced by the same configured system that emits the target decision, so they can carry policy traces without being

literal copies of the label. This is why the no-application-context ablation, the strict proxy-minimal stress test, and the determinism audit matter; together they show that suppressing one overt proxy family is not enough to make the task proxy-free, and that even aggressive proxy removal leaves a measurable decision surface. The result aligns with broader warnings from cross-dataset, lifecycle-aware, and shortcut-detection studies that closed-world performance can reflect data provenance, split design, operational regularities, or shortcut structure rather than general security reasoning [6–8].

The model-free audit separates information structure from learner performance. Compact field combinations determine the label within the observed export, but 64.70–71.35% of rows occupy singleton key contexts. Agreement persists at higher support thresholds, all seven full-export keys reappear in training-only discovery, and a split-spurious eighth candidate is rejected by one test error. Cardinality-preserving controls also break exactness, although by only 7–59 errors. The result is therefore empirical and noncausal; the keys summarize export-specific decision traces rather than the rule order, policy objects, implicit defaults, or device internals that generated them [4, 65].

Uncertainty diagnostics show why near-deterministic does not mean uncertainty-free. In random and temporal core settings, entropy, Brier score, and ECE remain extremely low, but uncertainty rises under feature restriction and some context-held-out pipelines. Probability-threshold conformal sets, APS expansion, and selective risk connect the firewall case to approximate-reasoning questions about validity, imprecision, and task-dependent uncertainty measurement [46, 48–50]. These diagnostics overlap through fitted probabilities and are complementary views, not independent replications. The rare-Deny Mondrian result shows how marginal validity can hide class-conditional undercoverage [35, 36, 39]. Under application/category shift, native CatBoost recovers Deny F1 from 0.0328 to 0.9774 relative to ordinal LightGBM, yet probability-threshold sets still collapse toward empty outputs and temperature scaling changes fixed-threshold queue capture in opposite directions across the two models. Uncertainty and simulated review behavior are therefore representation- and calibration-dependent; no context-shift coverage guarantee or model-invariant queue rule is claimed [52].

The context-held-out tests provide a stronger warning than the random and time-ordered splits, but not a single model-invariant ranking of context difficulty. Random hold-outs and within-window temporal checks indicate stability inside the observed export. Ordinal-LightGBM group hold-outs expose fragility when familiar contexts are withheld, and five-partition reruns show that the contrast across context definitions is not partition-invariant. On identical seed-42 rows, however, native-categorical CatBoost raises application/category macro-F1 from the 0.673–0.686 ordinal range to 0.992, while it does not improve destination-service hold-out. The boundary therefore combines context novelty, category OOV, representation, and learner behavior. This pattern is consistent with cross-evaluation and drift-aware IDS studies in which performance can degrade under dataset shift, concept drift, federation, or unseen environments [6, 57, 66]. Unfamiliar firewall contexts may be evidence-relative unknowns for a deployed pipeline even when they are not intrinsically noisy [52, 67].

Operationally, the most defensible use cases are policy-centered. A trained model may help flag unusual context, data-quality problems, or configuration changes for review, especially when combined with explainable log analytics or feature inspection [17, 26, 30]. It cannot replace deterministic rule analysis, policy ownership, or change control and remains complementary to rule-graph policy analysis [20–22]. Current logic-driven anomaly detection returns formally checkable findings from the rule set [11], and policy-configuration decision support ranks the resulting anomalies for the responsible managers [12], whereas the diagnostics reported here summarize observed decisions and carry empirical uncertainty. Confidence, conformal abstention, APS expansion, unfamiliar contexts, and residual pockets are simulated diagnostic signals whose operating points require model-specific calibration and validation against real review outcomes.

As a concrete instance of this operational reading, the rule-held-out diagnostic provides a within-export analogue of a configuration-change rehearsal. When whole rule contexts are withheld from training, macro-F1 falls from the near-deterministic core to 0.9275 (Table 14), the selective review queue at confidence 0.999 grows to 2.54% of records at selective risk 0.0025, and the dominant residual pocket is a single Allow-versus-Deny interaction, not diffuse error (Table 19). A network-management team could treat this signature as a prompt to inspect whether a rule context, representation, or logging condition changed. This remains a within-export decision-consistency demonstration; the study does not validate the queue against rule-graph ground truth, expert audit outcomes, or logged policy-change events.

The scalability evidence supports this operational reading but should not be overstated. Million-row reconstruction on modest cloud hardware shows that the protocol is feasible for practical log-analysis workflows, and the use of tree ensembles is consistent with broader tabular-security evidence that such models remain strong, efficient, and interpretable baselines [29]. At the same time, runtime feasibility

does not solve the harder validation problem. Deployment-oriented security analytics must still document data provenance, preprocessing boundaries, class imbalance, calibration, uncertainty, and drift behavior before treating high scores as evidence of operational relevance [32, 34].

Several limitations bound these claims. The target is an operational firewall decision, not an independently adjudicated security state, so Drop and Deny are not inherently malicious and Allow is not forensically benign [32, 68]. Label-source exclusion, training-only preprocessing, and proxy ablations reduce but cannot remove indirect proxies, which are the phenomenon under study and not merely a nuisance. The determinism audit is exact for this export but could be falsified by additional records. The file is a tool-capped export from one enterprise environment, one logging configuration, one firewall software configuration, and an observed 46-minute and 34-second timestamp span; it is not evidence for 24-hour, weekday, seasonal, or policy-cycle representativeness. The schema-independent components are instantiated on one independent public firewall dataset (Section 5.6), but this is not external validation of the enterprise policy-conditioned layer. Cross-organization, vendor, policy-regime, long-horizon drift, policy-rule-graph, sequence-modeling, and deployment-feedback validation remain outside scope and are required before that layer is generalized [6, 57, 66].

7 Conclusion

Firewall decisions in this single Palo Alto traffic-log export are redundantly and almost entirely reconstructable from structured operational context, even after direct label-source fields and overt application-context proxies are removed. A stricter proxy-minimal XGBoost stress test remains substantially reconstructable but no longer reaches the numerical ceiling, while LightGBM instability prevents a model-independent magnitude claim. The model-free audit finds seven minimum four-field determining sets in the all-field view, all using Source Port and session-end Bytes. Repeated-context and train-only checks retain exact seen-context agreement, but limited coverage and cardinality controls keep these sets export-specific and noncausal; trained models only show that the remaining context makes this label attainable at scale. Parallel downstream diagnostics confirm that conformal-set behavior and fixed-threshold queues depend on representation and calibration, not context novelty alone. Near-perfect firewall-action prediction is evidence about the observed decision process, not a claim about attack detection, causal policy logic, or cross-environment security generalization.

Future work should test the same framing across time-separated exports, organizations, vendors, longer validation windows, policy-rule representations, and deployment settings where review queues can be evaluated against real policy-change workflows. Stronger external validation would require repeated exports across policy revisions, richer rule-graph information, and controlled comparisons between pre-decision context, post-decision telemetry, and explicitly policy-proximal fields. Future studies should also examine whether conformal, selective, and drift-aware diagnostics can support operational review without creating a false sense of automated adjudication. For network-management analytics, the practical lesson is that firewall-log studies should report label construction, excluded fields, proxy risks, imbalance handling, ablations, per-class behavior, calibration, uncertainty, runtime environment, and reproducibility artifacts before drawing security claims from very high scores.

Statements and Declarations

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval and consent to participate No human participants were recruited and no animal data were used. The processed analysis table contains operational firewall telemetry and excludes source or destination IP addresses and user identifiers. It was derived from logs obtained under written authorization from the administration of the operating institution, which is not named here to protect its operational security posture. No ethics committee approval was required for analysis of this authorized, identifier-excluded operational table. This statement is limited to the processed analysis table and does not characterize every field that may have existed in the source logging environment. Raw and record-level processed logs are not publicly redistributed.

Consent for publication Not applicable.

Data availability Code, scripts, aggregate outputs, diagnostic summaries, figures, citation metadata, and pinned requirements are available at https://github.com/canay/traffic-scale-decision. Raw and record-level processed firewall logs are not shared because institutional authorization does not permit redistribution of policy-proximal operational records. Full reruns require authorized access to the same or an equivalently approved dataset; the public materials support audit of the reported aggregate results.

Author contributions Özkan Canay: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. Cem Özkurt: Conceptualization, Methodology, Validation, Formal analysis, Supervision, Writing – review & editing. Fatmir Garri: Data curation, Investigation, Software, Validation, Writing – review & editing.

Generative AI and AI-assisted technologies During the preparation of this work, the authors used OpenAI ChatGPT/Codex and Anthropic Claude/Code for English translation, language editing, manuscript-structure review, reference-format checking, Python code debugging, assistance with planning and executing scripted experiments, and review of aggregate outputs for manuscript integration. All experimental designs, code, configurations, outputs, interpretations, and manuscript claims were checked and approved by the authors, who take full responsibility for the content of the article.

References

[1] Abu Al-Haija, Q., Ishtaiwi, A.: Machine learning based model to identify firewall decisions to improve cyber-defense. International Journal on Advanced Science, Engineering and Information Technology **11**(4), 1688 (2021) https://doi.org/10.18517/ijaseit.11.4.14608

[2] Aljabri, M., Alahmadi, A.A., Mohammad, R.M.A., Aboulnour, M., Alomari, D.M., Almotiri, S.H.: Classification of firewall log data using multiclass machine learning models. Electronics **11**(12), 1851 (2022) https://doi.org/10.3390/electronics11121851

[3] Komadina, A., Kovačević, I., Štengl, B., Groš, S.: Comparative analysis of anomaly detection approaches in firewall logs: Integrating light-weight synthesis of security logs and artificially generated attack detection. Sensors **24**(8), 2636 (2024) https://doi.org/10.3390/s24082636

[4] Madhloom, J.K., Noori, Z.H., Ebis, S.K., Hassen, O.A., Darwish, S.M.: An information security engineering framework for modeling packet filtering firewall using neutrosophic Petri nets. Computers **12**(10), 202 (2023) https://doi.org/10.3390/computers12100202

[5] Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. Patterns **4**(9), 100804 (2023) https://doi.org/10.1016/j.patter.2023.100804

[6] Apruzzese, G., Pajola, L., Conti, M.: The cross-evaluation of machine learning-based network intrusion detection systems. IEEE Transactions on Network and Service Management **19**, 5152–5169 (2022) https://doi.org/10.1109/TNSM.2022.3157344

[7] Sudyana, D., Lin, Y.-D., Verkerken, M., Hwang, R.-H., Lai, Y.-C., D'Hooge, L., Wauters, T., Volckaert, B., De Turck, F.: Improving generalization of ML-based IDS with lifecycle-based dataset, auto-learning features, and deep learning. IEEE Transactions on Machine Learning in Communications and Networking **2**, 645–662 (2024) https://doi.org/10.1109/TMLCN.2024.3402158

[8] Müller, N.M., Roschmann, S., Khan, S., Sperl, P., Böttinger, K.: Shortcut detection with variational autoencoders. In: 2024 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2024). https://doi.org/10.1109/IJCNN60899.2024.10650671

[9] Sánchez, L., Costa, N., Otero, J., Couso, I.: Physics-informed learning under epistemic uncertainty with an application to system health modeling. International Journal of Approximate Reasoning **161**, 108988 (2023) https://doi.org/10.1016/j.ijar.2023.108988

[10] Nife, F.N., Kotulski, Z.: Application-aware firewall mechanism for software defined networks. Journal of Network and Systems Management **28**(3), 605–626 (2020) https://doi.org/10.1007/s10922-020-09518-z

[11] Elfaki, A.O., Albluwi, A., Aljaedi, A., Nerma, M.H.M.: Explainable logic-driven firewall anomaly detection with knowledge graph visualization and machine learning validation. Electronics **15**(8), 1714 (2026) https://doi.org/10.3390/electronics15081714

[12] Park, J., Park, B., Kim, T.-S.: Development of an anomaly classification model and a decision support tool for firewall policy configuration. Applied Sciences **15**(6), 2979 (2025) https://doi.org/10.3390/app15062979

[13] Yılmaz, A., Das, R.: A novel hybrid approach combining GCN and GAT for effective anomaly detection from firewall logs in campus networks. Computer Networks **259**, 111082 (2025) https://doi.org/10.1016/j.comnet.2025.111082

[14] Rahman, M.H., Islam, T., Rana, M., Tasnim, R., Mona, T., Sakib, M.: Machine learning approach on multiclass classification of internet firewall log files. In: 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), pp. 358–364 (2023). https://doi.org/10.1109/CISES58720.2023.10183601

[15] Al-Behadili, H.N.K.: Decision tree for multiclass classification of firewall access. International Journal of Intelligent Engineering and Systems **14**(3), 294–302 (2021) https://doi.org/10.22266/ijies2021.0630.25

[16] Efeoglu, E., Tuna, G.: Classification of firewall log files with different algorithms and performance analysis of these algorithms. Journal of Web Engineering **23**, 561–594 (2024) https://doi.org/10.13052/jwe1540-9589.2344

[17] Fatima, Z., Hussain, R., Dilshad, A., Shakir, M., Laghari, A.: Explainable AI-driven firewall evaluation: Empowering cybersecurity decision-making for optimal network defense. Security and Privacy **8**(3) (2025) https://doi.org/10.1002/spy2.70026

[18] Bairwa, A.K., Kamboj, A., Joshi, S., Pljonkin, A.P., Hiranwal, S.: Prediction of TCP firewall action using different machine learning models. In: Communications in Computer and Information Science, pp. 161–174 (2024). https://doi.org/10.1007/978-3-031-70807-7_12

[19] Korkmaz, A., Bulut, S., Talan, T., Kosunalp, S., Iliev, T.: Enhancing firewall packet classification through artificial neural networks and synthetic minority over-sampling technique: An innovative approach with evaluative comparison. Applied Sciences **14**(16), 7426 (2024) https://doi.org/10.3390/app14167426

[20] Al-Shaer, E.S., Hamed, H.H.: Modeling and management of firewall policies. IEEE Transactions on Network and Service Management **1**(1), 2–10 (2004) https://doi.org/10.1109/TNSM.2004.4623689

[21] Yuan, L., Chen, H., Mai, J., Chuah, C.-N., Su, Z., Mohapatra, P.: FIREMAN: A toolkit for firewall modeling and analysis. In: 2006 IEEE Symposium on Security and Privacy (S&P'06), pp. 199–213 (2006). https://doi.org/10.1109/SP.2006.16

[22] Hu, H., Ahn, G.-J., Kulkarni, K.: Detecting and resolving firewall policy anomalies. IEEE Transactions on Dependable and Secure Computing **9**(3), 318–331 (2012) https://doi.org/10.1109/TDSC.2012.20

[23] Huhtala, Y., Kärkkäinen, J., Porkka, P., Toivonen, H.: TANE: An efficient algorithm for discovering functional and approximate dependencies. The Computer Journal **42**(2), 100–111 (1999) https://doi.org/10.1093/comjnl/42.2.100

[24] Papenbrock, T., Naumann, F.: A hybrid approach to functional dependency discovery. In: Proceedings of the 2016 International Conference on Management of Data, pp. 821–833. ACM, San Francisco, CA, USA (2016). https://doi.org/10.1145/2882903.2915203

[25] Talukder, M., Islam, M., Uddin, M., Hasan, K., Sharmin, S., Alyami, S., Moni, M.: Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. Journal of Big Data **11**(1), 33 (2024) https://doi.org/10.1186/s40537-024-00886-w

[26] Le, T., Kim, H., Kang, H., Kim, H.: Classification and explanation for intrusion detection system based on ensemble trees and SHAP method. Sensors **22**(3), 1154 (2022) https://doi.org/10.3390/s22031154

[27] Aouedi, O., Piamrat, K., Parrein, B.: Ensemble-based deep learning model for network traffic classification. IEEE Transactions on Network and Service Management **19**, 4124–4135 (2022) https://doi.org/10.1109/TNSM.2022.3193748

[28] Rehman, A., Saba, T., Jamjoom, M.M., Al-Otaibi, S., Khan, M.I.: Advances in machine learning for explainable intrusion detection using imbalance datasets in cybersecurity with Harris hawks optimization. Computers, Materials & Continua **86**(1), 1–15 (2026) https://doi.org/10.32604/cmc.2025.068958

[29] Zoppi, T., Gazzini, S., Ceccarelli, A.: Anomaly-based error and intrusion detection in tabular data: No DNN outperforms tree-based classifiers. Future Generation Computer Systems **160**, 951–965 (2024) https://doi.org/10.1016/j.future.2024.06.051

[30] Alam, K., Kifayat, K., Sampedro, G., Karovic, V., Naeem, T.: SXAD: Shapely eXplainable AI-based anomaly detection using log data. IEEE Access **12**, 95659–95672 (2024) https://doi.org/10.1109/access.2024.3425472

[31] Devabhaktuni, H., Nair, S.S., C, T., Mohan, A., Yeruva, N.M.: Real-time Windows event log analysis and anomaly detection using explainable machine learning. In: 2025 1st IEEE Uttar Pradesh Section Women in

Engineering International Conference on Electrical Electronics and Computer Engineering (UPWIECON), pp. 448–454 (2025). https://doi.org/10.1109/UPWIECON67212.2025.11390185

[32] Anis, F., Alabdullatif, M., Aljbli, S., Hammoudeh, M.: A survey on the applications of deep learning in network intrusion detection systems to enhance network security. IEEE Access **13**, 185357–185373 (2025) https://doi.org/10.1109/ACCESS.2025.3624952

[33] Verkerken, M., D'hooge, L., Volckaert, B., De Turck, F., Apruzzese, G.: "What is the problem space?" Defining host-space adversarial perturbations against network intrusion detection systems. In: Proceedings of the ACM Asia Conference on Computer and Communications Security, pp. 1043–1059 (2026). https://doi.org/10.1145/3779208.3807482

[34] Youssef, M.A.: CALIBURN: Operationally calibrated streaming intrusion detection with regime-dependent conformal risk control. arXiv preprint arXiv:2605.24696 (2026) https://doi.org/10.48550/arxiv.2605.24696

[35] Bulavas, V., Marcinkevicius, V., Ruminski, J.: Study of multi-class classification algorithms' performance on highly imbalanced network intrusion datasets. Informatica **32**, 441–475 (2021) https://doi.org/10.15388/21-INFOR457

[36] Bagui, S., Li, K.: Resampling imbalanced data for network intrusion detection datasets. Journal of Big Data **8**(1), 6 (2021) https://doi.org/10.1186/s40537-020-00390-x

[37] Maseer, Z., Yusof, R., Bahaman, N., Mostafa, S., Foozy, C.: Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset. IEEE Access **9**, 22351–22370 (2021) https://doi.org/10.1109/ACCESS.2021.3056614

[38] Song, J., Wang, X., He, M., Jin, L.: CSK-CNN: Network intrusion detection model based on two-layer convolution neural network for handling imbalanced dataset. Information **14**(2), 130 (2023) https://doi.org/10.3390/info14020130

[39] Farhadpour, S., Warner, T., Maxwell, A.: Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices. Remote Sensing **16**(3), 533 (2024) https://doi.org/10.3390/rs16030533

[40] Apicella, A., Isgro, F., Prevete, R.: Don't push the button! Exploring data leakage risks in machine learning and transfer learning. Artificial Intelligence Review **58**(11), 339 (2025) https://doi.org/10.1007/s10462-025-11326-3

[41] Chou, D., Jiang, M.: A survey on data-driven network intrusion detection. ACM Computing Surveys **54**, 1–36 (2021) https://doi.org/10.1145/3472753

[42] Arcos-Argudo, M., Bojorque, R., Torres, A.: A deterministic comparison of classical machine learning and hybrid deep representation models for intrusion detection on NSL-KDD and CICIDS2017. Algorithms **18**(12), 749 (2025) https://doi.org/10.3390/a18120749

[43] Harsapranata, A.I., Sediyono, E., Purnomo, H.D.: Intrusion detection: Boruta feature selection and semi-supervised outlier clustering with multi-dataset evaluation. Engineering, Technology & Applied Science Research **15**(6), 30283–30289 (2025) https://doi.org/10.48084/etasr.14351

[44] Xin, C., Xu, K.: Cross-dataset Transformer-IDS with calibration and AUC optimization (evaluated on NSL-KDD, UNSW-NB15, CIC-IDS2017). Journal of Cyber Security **7**(1), 483–503 (2025) https://doi.org/10.32604/jcs.2025.071627

[45] Tawfik, M., Badawy, W.: Explainable deep learning for Modbus/TCP traffic classification in industrial control systems. In: 2025 1st Future International Conference on Artificial Intelligence and Cybersecurity (FICAC), pp. 111–118 (2025). https://doi.org/10.1109/FICAC65757.2025.11341864

[46] Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York, NY (2005). https://doi.org/10.1007/b106715

[47] Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning **16**(4), 494–591 (2023) https://doi.org/10.1561/2200000101

[48] Romano, Y., Sesia, M., Candès, E.: Classification with valid and adaptive coverage. In: Advances in Neural Information Processing Systems, vol. 33, pp. 3581–3591 (2020). https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html

[49] Cella, L., Martin, R.: Validity, consonant plausibility measures, and conformal prediction. International Journal of Approximate Reasoning **141**, 110–130 (2022) https://doi.org/10.1016/j.ijar.2021.07.013

[50] Haas, S., Hüllermeier, E.: Uncertainty quantification in ordinal classification: A comparison of measures. International Journal of Approximate Reasoning **186**, 109479 (2025) https://doi.org/10.1016/j.ijar.2025.

109479

[51] Chow, C.K.: On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory **16**(1), 41–46 (1970) https://doi.org/10.1109/TIT.1970.1054406

[52] Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning **110**(3), 457–506 (2021) https://doi.org/10.1007/s10994-021-05946-3

[53] Sadinle, M., Lei, J., Wasserman, L.: Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association **114**(525), 223–234 (2019) https://doi.org/10.1080/01621459.2017.1395341

[54] Gupta, P., Wever, M., Hüllermeier, E.: Information leakage detection through approximate Bayes-optimal prediction. Information Sciences **719**, 122419 (2025) https://doi.org/10.1016/j.ins.2025.122419

[55] Yang, S., Zheng, X., Li, J., Xu, J., Zhang, X., Ngai, E.: Self-supervised adaptation method to concept drift for network intrusion detection. IEEE Transactions on Dependable and Secure Computing **22**, 7632–7646 (2025) https://doi.org/10.1109/TDSC.2025.3599321

[56] Shaker, M.H., Hüllermeier, E.: Random forest calibration. Knowledge-Based Systems **328**, 114143 (2025) https://doi.org/10.1016/j.knosys.2025.114143

[57] Bilal, M.A., Ul Islam, I., Idrees, S., Qasim, M., Khan, M.J., Khan, J.: Dataset-centric evaluation of federated intrusion detection models in IoT networks. Scientific Reports **16**, 2683 (2026) https://doi.org/10.1038/s41598-025-32567-w

[58] Ayodele, B., Buttigieg, V.: Evaluating cross-domain generalisation of deep learning models in network traffic classification: A statistical and experimental analysis of VNF and PNF. Journal of Network and Systems Management **34**(3), 73 (2026) https://doi.org/10.1007/s10922-026-10049-2

[59] Botacin, M., Gomes, H.M.: Towards more realistic evaluations: The impact of label delays in malware detection pipelines. Computers & Security **148**, 104122 (2025) https://doi.org/10.1016/j.cose.2024.104122

[60] Nam, S., Jeong, E.-D., Hong, J.W.-K.: Log-TF-IDF and NETCONF-based network switch anomaly detection. International Journal of Network Management **35**(1) (2025) https://doi.org/10.1002/nem.2322

[61] Palo Alto Networks: Session Logging Considerations. Accessed 12 July 2026 (n.d.). https://docs.paloaltonetworks.com/network-security/security-policy/administration/security-rules/session-logging-considerations

[62] Cotton, M., Eggert, L., Touch, J., Westerlund, M., Cheshire, S.: Internet assigned numbers authority (IANA) procedures for the management of the service name and transport protocol port number registry. RFC 6335, RFC Editor (2011). https://doi.org/10.17487/RFC6335

[63] UCI Machine Learning Repository: Internet Firewall Data. CC BY 4.0 (2019). https://doi.org/10.24432/C5131M

[64] Hindawi, M., Bodair, H., Abdel-Nabi, H.: Firewall action classification enhanced by machine learning: A study on internet traffic logs. In: 2025 26th International Arab Conference on Information Technology (ACIT), pp. 814–819 (2025). https://doi.org/10.1109/ACIT68900.2025.11510602

[65] Lee, J.-K., Hong, T., Lee, G.: AI-based approach to firewall rule refinement on high-performance computing service network. Applied Sciences **14**(11), 4373 (2024) https://doi.org/10.3390/app14114373

[66] Yang, S., Zheng, X., Li, J., Xu, J., Wang, X., Ngai, E.: ReCDA: Concept drift adaptation with representation enhancement for network intrusion detection. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3818–3828 (2024). https://doi.org/10.1145/3637528.3672007

[67] Jorgensen, S., Holodnak, J., Dempsey, J., De Souza, K., Raghunath, A., Rivet, V., Demoes, N., Alejos, A., Wollaber, A.: Extensible machine learning for encrypted network traffic application labeling via uncertainty quantification. IEEE Transactions on Artificial Intelligence **5**, 420–433 (2024) https://doi.org/10.1109/TAI.2023.3244168

[68] Ahsan, M., Nygard, K., Gomes, R., Chowdhury, M., Rifat, N., Connolly, J.: Cybersecurity threats and their mitigation approaches using machine learning: A review. Journal of Cybersecurity and Privacy **2**, 527–555 (2022) https://doi.org/10.3390/jcp2030027