

Design and Evaluation of an Explainable Network-Edge IDPS for Early Detection and Reversible Blocking of Malicious DNS-over-HTTPS Tunnels

Kurt Louise T. Dacung, Jan Viktor V. Esparrago, James Ethan De Jesus, Jasper Lincolnd P. Flores, Sem T. Borja

College of Computing and Information Sciences (BS Information Technology, Major in Network Administration), Systems Plus College Foundation, Philippines

Corresponding Author (Kurt Louise Dacung) Email: kurtlouisedacung1@gmail.com

ABSTRACT

DNS over HTTPS (DoH) protects name-resolution traffic from passive inspection, but the same encryption can conceal tunneling behavior from controls that rely on visible DNS content. This proposal presents an explainable network-edge intrusion detection and prevention system (IDPS) for a commodity Linux router. The prototype will observe only timing, direction, and size features from the first N payload-bearing TCP segments of an encrypted connection; it will not decrypt TLS, inspect DNS questions, or identify individual HTTP/2 streams. A class-weighted logistic regression baseline and a bounded random forest will be developed using group-separated fitting, model selection, calibration, threshold tuning, safety confirmation, and final testing. The primary corpus will be the raw CIRA-CIC-DoHBrw-2020 packet captures, while DoH-Tunnel-Traffic-HKD will serve as a locked positive-only challenge. The selected model will be compared with routing-only, pinned Suricata-only, and hybrid configurations at registered offered loads. Evaluation will cover precision-recall performance, calibration, false blocks, verdict coverage, pre-verdict exposure, latency, throughput, packet loss, CPU use, and memory use. Temporary exact-connection blocking will be enabled only if the adviser approves the eight-week route and the benign-risk confidence bound meets the pre-registered safety gate. The expected output is a reproducible evidence package showing whether early learned detection adds safe, measurable value to a conventional firewall and rule-based IPS on low-cost hardware.

Keywords: *DNS over HTTPS, intrusion detection and prevention system, encrypted traffic analysis, explainable artificial intelligence, Suricata, network-edge security*

SECTION I — INTRODUCTION (THE PROBLEM AND ITS SETTING)

1.1 Introduction

Domain Name System (DNS) resolution is necessary before most Internet services can be reached. Traditional DNS exposes query names in clear text, while DNS over HTTPS carries the exchange inside encrypted HTTPS. RFC 8484 improves privacy between a client and a resolver, but it also removes content that network administrators commonly use to distinguish ordinary resolution from covert command, control, or data transfer (Hoffman & McManus, 2018). A normal firewall can allow or deny an address, port, protocol, or known connection state; it does not by itself determine whether two permitted encrypted connections behave differently. This distinction explains why the proposed prototype includes a firewall, Suricata, and a separate learned traffic monitor rather than treating the components as substitutes.

The study will design and evaluate a Linux router that observes the early behavior of forwarded, TCP-based DoH connections. Its learned detector will summarize the first N payload-bearing TCP segments by direction, size, timing, and burst behavior. The inference unit is an aggregate connection prefix—not an individual DNS query—because HTTP/2 can interleave multiple streams within one TCP connection (Thomson & Benfield, 2022). When a calibrated score crosses a frozen threshold, the system will record an alert and, only after both approval and safety conditions are satisfied, add that exact bidirectional IPv4 five-tuple to an expiring nftables set. The limited scope avoids a claim that encrypted content has been recovered or that the system can identify the specific domain or file carried in the tunnel.

The proposal addresses an applied network-administration question: can early learned evidence add useful detection to a pinned rule-based IPS without imposing an unacceptable availability risk? Prior work has already established that DoH traffic can be classified and explained (MontazeriShatoori et al., 2020; Zebin et al., 2022), and recent systems study real-time or high-speed detection (Du et al., 2026). The intended contribution is therefore not a new machine-learning algorithm. It is a reproducible enforcement experiment that freezes the decision point, prevents identity leakage, separates all data roles, measures false blocks with an exact confidence bound, and compares the learned policy against Suricata on commodity hardware.

Research Questions

1. At a validation-selected and then frozen packet budget N , how accurately and how well calibrated can malicious DoH tunnel connection prefixes be distinguished from benign DoH and ordinary HTTPS without identity fields, TLS decryption, or a per-query assumption?
2. How does detection performance vary by benign traffic stratum, tunnel tool, resolver, and capture group, and what recall and verdict coverage are retained on the locked DoH-Tunnel-Traffic-HKD unseen-tool challenge set?
3. What incremental detection benefit and operational cost does the learned policy add relative to routing-only, Suricata-only, and ML-only configurations?
4. Does the automatic-blocking policy satisfy its pre-registered benign-risk gate, and do timeout, rollback, controller-stop, and Suricata-bypass tests restore or preserve traffic as specified?

1.2 Background of the Study

DoH was standardized to reduce surveillance and manipulation of DNS exchanges. Its use of HTTPS allows resolution traffic to share the same port and transport security mechanisms as ordinary web traffic. Böttger et al. (2019) found that this design can provide privacy with limited effect on page-load performance, helping explain its practical adoption. At the same time, encrypted DNS remains susceptible to traffic analysis: packet direction, size, timing, and burst structure can reveal patterns even when content is unavailable (Siby et al., 2020). These observable side channels create a narrow opportunity for network monitoring, but they do not justify reading meaning into a single packet or claiming knowledge of encrypted payload contents.

CIRA-CIC-DoHBrw-2020 is frequently used to study malicious DoH. Its official inventory includes browser-generated HTTPS and benign DoH traffic and tunnels produced by dns2tcp, DNSCat2, and Iodine across several public resolvers, with raw packet captures made available for alternative feature extraction (Canadian Institute for Cybersecurity, 2020). Published classifiers report strong results on this corpus, including time-series methods, explainable models, ensemble methods, and lightweight two-stage schemes (Abu Al-Haija et al., 2023; Aggarwal & Kumar, 2024; Singh & Roy, 2022; Zebin et al., 2022). These results establish feasibility, but they do not automatically establish safe live blocking. Flow tables may include final duration or total-byte information unavailable at an early decision, and random row splits can place closely related flows from the same capture scenario in both training and test sets. Such leakage can produce optimistic performance estimates (Kapoor & Narayanan, 2023).

The official CIRA inventory also cannot be treated as the exact sample count for this study. The published browser total is 917,300 flows; the open combined release separates 897,493 non-DoH rows and 19,807 normal-DoH rows. The official tunnel table totals 249,750 flows, whereas the combined release contains 249,836 CIRA-origin tunnel-tool rows. These figures motivate a hard pre-experiment audit of raw files, hashes, labels, duplicates, capture groups, and first- N eligibility rather than copying a headline count into the statistical analysis. Yusof et al. (2022) likewise emphasize the value of visual and distributional examination of the dataset before modeling.

A second challenge is the difference between detection and prevention. An offline classifier can make an error without interrupting service. An inline system may break legitimate name resolution, web access, or administrative activity. NIST describes an IDPS as a combination of monitoring, analysis, and response capabilities, but its guidance also makes deployment, tuning, and operational impact part of system evaluation (Scarfone & Mell, 2007). The present study therefore treats automatic blocking as a separately qualified action. Probability calibration, a held-out threshold, allowlisting, expiring blocks, rollback, and fail-open behavior are not add-ons; they are the conditions under which an alert may become an enforcement action.

The project is also aligned with the Philippine National Cybersecurity Plan 2023–2028, adopted through Executive Order No. 58, which calls for more resilient cyber defenses and stronger cybersecurity capacity (Office of the President of the Philippines, 2024). The policy establishes national relevance but does not prove that local incidents are caused by DoH tunnels. Accordingly, this proposal makes no prevalence claim. It evaluates a contained laboratory prototype that could inform network-administration practice in resource-constrained educational or small-organization settings.

Statement of the Research Gap

The gap is not the absence of artificial intelligence, explainability, or real-time DoH detection. The focal gap is the lack, among the reviewed sources, of a single commodity-gateway study that combines all of the following under one pre-registered evaluation: (a) a frozen first- N TCP-connection-prefix decision valid under HTTP/2 multiplexing; (b) identity-free and leakage-resistant predictors; (c) group-separated fitting, selection, calibration, threshold tuning, safety confirmation, and final testing; (d) temporary exact-connection blocking with expiration, rollback, and fail-open controls; and (e) a four-mode comparison with a pinned Suricata IPS that jointly measures detection quality, false blocks, exposure before verdict, latency, throughput, packet loss, CPU use, and memory use. The research target is therefore evidence about safe incremental enforcement on a low-cost edge router—not another claim of high offline accuracy.

1.3 Objectives of the Study

General Objective

The general objective is to design and evaluate an explainable network-edge intrusion detection and prevention system for early detection of malicious TCP-based DNS-over-HTTPS tunnels and cautious, reversible response on commodity hardware without decrypting TLS.

Specific Objectives

1. Construct a physically and logically contained Linux routed-gateway testbed that integrates nftables, Suricata 8.0.1, a passive early-flow machine-learning monitor, and an expiring exact-connection block controller.
2. Audit, hash, label, group, and extract eligible connection prefixes from the raw CIRA-CIC-DoHBrw-2020 packet captures and preserve DoH-Tunnel-Traffic-HKD as a locked positive-only external challenge.
3. Develop leakage-resistant feature vectors at candidate budgets $N = 20, 40, \text{ and } 60$ payload-bearing TCP segments, excluding identity fields, post-decision data, TLS content, and per-query assumptions.
4. Fit and compare a class-weighted logistic regression baseline and a bounded random forest; select one packet budget and model on grouped development data, then calibrate probabilities and freeze the action threshold on separate groups.
5. Determine whether the frozen policy satisfies a two-sided 95% exact benign-risk upper bound of 0.0001 separately for ordinary HTTPS and benign DoH before enabling automatic blocking.
6. Implement asynchronous SHAP-based feature attribution for audit records without placing explanation computation in the enforcement path.
7. Compare routing-only, Suricata-only, ML-only, and hybrid configurations under controlled offered loads while measuring detection benefit, false blocks, verdict coverage, pre-verdict exposure, latency, throughput, packet loss, CPU use, and memory use.
8. Verify timed expiration, manual rollback, monitor-stop behavior, Suricata-bypass behavior, logging integrity, and reproducibility of the final evidence package.

1.4 Scope and Limitation of the Study

Scope

- The prototype will operate as an IPv4 layer-3 Linux router between an isolated generator and sink. It will evaluate TCP-based DoH carried over TLS and HTTP/2.
- The learned decision will use only the first selected N payload-bearing TCP segments after connection establishment. Pure acknowledgments and retransmissions will not advance the packet budget.
- Permitted predictors are per-direction packet and byte counts; payload-length minimum, maximum, mean, and standard deviation; inter-arrival mean and standard deviation; direction changes; directional burst counts, sizes, and durations; and explicitly defined upload/download ratios with zero handling.
- The study will compare two bounded models: class-weighted logistic regression and random forest. It will use Platt or isotonic calibration selected on held-out development groups and SHAP only for post-action feature attribution.
- The main dataset will be CIRA-CIC-DoHBrw-2020 raw PCAP. The HKD corpus will be a locked challenge for recall and coverage only. A prospectively generated benign-DoH safety corpus may supplement the safety-confirmation set but may not be used for fitting, model selection, calibration, or threshold selection.
- The operational benchmark will use four modes—routing only, Suricata IPS, ML policy, and hybrid—at 10, 50, and 100 Mbps, or at a lower frozen grid if routing-only preflight shows saturation. Each mode-load combination will be repeated three times for ten minutes.
- The preferred route is an eight-week IDPS implementation subject to adviser approval. If approval is not granted or the safety gate is not met, the same prototype and evaluation will continue in alert-only IDS mode without presenting automatic prevention as completed.

Limitations and Delimitations

- The study will not decrypt TLS, recover DNS names, inspect message content, or identify which HTTP/2 stream or DNS query is malicious.
- HTTP/3, QUIC, IPv6, non-DoH DNS tunnels, endpoint telemetry, browser instrumentation, and production deployment are outside scope.
- Public corpora contain synthetic tunnel tools and historical client-resolver combinations. Findings will be bounded to the tested data, capture groups, loads, and hardware; they will not establish universal detection.
- The external HKD challenge has no benign class. It can support recall and verdict-coverage statements, but not precision, specificity, false-positive rate, or calibrated prevalence claims.
- SHAP values will describe model-feature attribution for the recorded prediction. They will not be interpreted as causal explanations, proof of analyst usefulness, or decrypted evidence.
- Replay repetitions will estimate operational variability but will not be counted as new independent classification samples. Load will be created through concurrent time-preserved flows rather than global capture acceleration, because acceleration would alter timing features.
- No live malware, personal communication, campus production traffic, human participants, or Internet-routed malicious traffic will be used.

1.5 Significance of the Study

Table 1 lists the intended beneficiaries and the bounded value each is expected to receive from the study.

Stakeholder	Expected value
Network administrators	The study provides a measurable basis for deciding whether learned early-flow evidence adds value beyond stateful firewalling and a signature-based IPS, including the cost of false blocks and resource use.

Small organizations and educational laboratories	The commodity-hardware design and contained replay protocol offer a reproducible path for evaluating encrypted-traffic defenses without purchasing a commercial appliance or inspecting private payload content.
Security researchers	The group-separated data roles, exact safety gate, frozen packet budget, and external positive-only challenge provide a stricter template for avoiding leakage and overstated generalization in encrypted-traffic classification.
Students and instructors	The prototype demonstrates the distinct functions of a firewall, IDS, IPS, calibrated machine-learning monitor, and reversible response controller in one defensible network-administration project.
Systems Plus College Foundation	The work can contribute a documented, isolated testbed and an evidence package for future teaching and research, without making unverified claims about the institution's production network.
Future researchers	The frozen configurations, hashes, group manifests, feature schema, test scripts, and results tables can support replication, extension to HTTP/3 or IPv6, and evaluation on more representative benign traffic.
Sustainable Development Goal 9	The project supports resilient infrastructure and responsible innovation through a low-cost, reproducible cybersecurity prototype. No broader SDG impact is claimed beyond this technical contribution (United Nations, n.d.).

Table 1. Intended beneficiaries and bounded contributions of the study.

1.6 Definition of Terms

Benign-risk gate. The pre-registered condition requiring the two-sided 95% Clopper–Pearson upper confidence limit for the false-block probability to be at most 0.0001, evaluated separately for eligible ordinary HTTPS and benign DoH prefixes.

Calibration. The agreement between a model's predicted probability and observed event frequency. Operationally, a held-out mapping such as Platt scaling or isotonic regression will convert raw scores into probabilities before thresholding.

Commodity hardware. General-purpose computers and network interfaces meeting the registered minimum specifications, rather than a specialized commercial security appliance.

Connection prefix. The observable portion of one bidirectional TCP connection from establishment through the selected first N payload-bearing segments. It is the study's unit of inference.

DNS over HTTPS (DoH). A standardized method for sending DNS queries and responses through HTTPS. In this study, only TCP-based DoH is evaluated.

DoH tunnel. A tested mechanism that carries non-DNS data or command traffic through DNS messages transported over HTTPS. The label comes from controlled corpus generation, not payload inspection by the prototype.

Explainable artificial intelligence. Methods that attribute a prediction to input features. Operationally, SHAP values will be generated asynchronously from stored feature vectors for audit, not for causal interpretation.

Fail-open. A designed behavior in which forwarded traffic continues when the learned monitor or Suricata inspection path becomes unavailable, while loss of inspection is logged.

False block. An eligible benign connection prefix for which the frozen policy causes an automatic temporary block.

Firewall. A rule-based packet or connection control. Here, nftables supplies allowlists, forwarding policy, NFQUEUE routing, and temporary exact-five-tuple block sets; it does not perform the learned classification.

Five-tuple. The source IP address, destination IP address, source port, destination port, and transport protocol that identify the tested IPv4 TCP connection.

Group separation. Assignment of all related observations from one source capture or registered scenario to one data role so that dependent rows cannot appear across fitting, selection, calibration, threshold, safety, or final-test partitions.

IDPS. An intrusion detection and prevention system that detects suspected malicious activity and may perform a response. Automatic prevention in this study remains conditional on approval and the safety gate.

Pre-verdict exposure. The number of packets, bytes, and elapsed time observed before the system reaches an eligible decision at the frozen packet budget.

Reversible blocking. A temporary action applied only to the exact connection and automatically removed by timeout, with a tested manual rollback path.

Suricata. The open-source network threat detection engine used as the pinned rule-based IPS comparator through Linux NFQUEUE.

Verdict coverage. The proportion of in-scope connections that reach the selected packet budget and therefore receive an ML verdict; short connections are reported separately rather than forced into a class.

SECTION II — LITERATURE REVIEW

2.1 Introduction to the Review

This review moves from global foundations to the Philippine national context, the local application setting, and focal technical studies. It first examines DoH standards, encrypted-traffic observability, and the role of IDPS controls. It then discusses Philippine policy and the bounded relevance of an educational commodity-gateway prototype. The focal portion compares datasets, machine-learning approaches, explainability, generalization, early decisions, and operational enforcement. Methodological literature on data leakage, calibration, class imbalance, exact confidence intervals, and reproducible data mining is integrated because the research gap depends as much on evaluation discipline as on classifier choice.

2.2 Review of Related Literature and Studies

2.2.1 Global Context: Encrypted DNS and Network Monitoring

RFC 8484 defines DoH as DNS exchanges carried through HTTPS and recommends HTTP/2 support (Hoffman & McManus, 2018). HTTP/2 allows multiple streams to share a connection and interleaves frames from those streams (Thomson & Benfield, 2022). This matters methodologically: packets from a TCP connection cannot be assumed to form one DNS request-response pair. The present study consequently predicts at the connection-prefix level and avoids language suggesting per-query localization. HTTP/3 carries HTTP semantics over QUIC rather than TCP, so its transport behavior differs and is explicitly excluded from the current prototype (Bishop, 2022).

DoH offers meaningful privacy, but encryption does not erase every observable characteristic. Böttger et al. (2019) documented the ecosystem and measured a limited page-load cost, supporting the practical value of encrypted DNS. Siby et al. (2020) demonstrated that timing and direction patterns can still support traffic analysis and that client, resolver, platform, and location changes reduce attack performance. Their results support two principles adopted here: traffic metadata can carry signal, but environmental variation must be treated as a generalization risk rather than hidden inside a random split.

Network controls serve different functions. NIST SP 800-94 distinguishes monitoring and analysis from prevention actions and stresses tuning, deployment location, and operational impact (Scarfone & Mell, 2007). A stateful firewall decides whether a packet or connection matches explicit policy; a signature IPS can recognize rule-described evidence; a learned monitor estimates similarity to patterns in its training task. Suricata's NFQUEUE deployment allows Linux-forwarded packets to be inspected inline, and the queue bypass option can preserve forwarding when the inspection application is absent (Open Information Security Foundation, 2026). The proposed layered comparison is therefore justified by complementary capabilities, not by the assumption that a firewall is unnecessary.

2.2.2 National Context: Philippine Cybersecurity and Responsible Monitoring

Executive Order No. 58 adopted the Philippine National Cybersecurity Plan 2023–2028 and directed its implementation. The plan's emphasis on resilient infrastructure, capacity building, and coordinated defense makes low-cost, evidence-based network security research nationally relevant (Office of the President of the Philippines,

2024). This policy, however, is not evidence of DoH-tunnel prevalence in Philippine institutions. It supports the need for defensible capability development while requiring the study to avoid claims that are not measured.

The Data Privacy Act of 2012 establishes principles for lawful and proportionate processing of personal information (Republic of the Philippines, 2012). Although the experiment will use public research corpora and controlled non-personal traffic rather than production communications, its architecture adopts data-minimizing choices: no payload decryption, no domain extraction, no identity fields as predictors, hashed connection identifiers in analysis outputs, and retention limited to reproducibility needs. These controls align the technical experiment with responsible monitoring practice.

2.2.3 Local Application Context

Within the sources reviewed for this proposal, no Philippine or institution-specific study was found that evaluates an early, calibrated DoH-tunnel blocking policy on a commodity routed gateway against a frozen Suricata baseline while measuring both false blocks and network-service cost. This statement is deliberately limited to the reviewed source set and is not a claim that no such implementation exists anywhere. The local contribution is therefore an application and evidence gap: translating established detection techniques into a contained, teachable, and auditable network-administration experiment at Systems Plus College Foundation.

The planned setting also affects the design. A capstone laboratory has limited time, hardware, and benign traffic volume. The project therefore uses two physical computers with virtualized generator and sink roles, freezes a manageable two-model search, and makes automatic blocking conditional. The design does not present the laboratory as a production network, nor does it collect student or employee traffic. This separation protects the study's internal validity and keeps its conclusions at the scale actually tested.

2.2.4 Focal Studies: DoH Classification, Tunnels, and Explainability

Vekshin et al. (2020) showed that extended flow information can identify DoH, while MontazeriShatoori et al. (2020) released the CIRA-CIC-DoHBrw-2020 task and used time-series features to detect DoH tunnels. Later studies applied conventional ensembles, two-stage feature reduction, and multi-class methods to the same benchmark (Abu Al-Haija et al., 2023; Aggarwal & Kumar, 2024; Casanova & Lin, 2023; Singh & Roy, 2022). These works demonstrate that encrypted traffic can be classified without decrypting payloads. They also show why another accuracy-only study would be weak: high benchmark scores already exist.

Zebin et al. (2022) integrated explainable AI into DoH attack classification, establishing that feature-level attribution is not a novel gap by itself. Pich et al. (2024) examined single-packet features, demonstrating interest in earlier evidence, while Du et al. (2026) proposed TunnelEye for explainable real-time detection in high-speed backbone settings. The present proposal does not compete on backbone scale or claim a new explanation method. Its distinction is the controlled connection-prefix decision, held-out calibration and safety qualification, exact temporary enforcement, and operational comparison on commodity hardware.

Mitsuhashi et al. (2022) studied persistent behavior for recognizing malicious tunnel tools and released the DoH-Tunnel-Traffic-HKD corpus containing dnstt, tcp-over-dns, and tuns traffic (Mitsuhashi et al., 2023). These tools differ from the three tunnel families in the CIRA corpus. That makes HKD useful as an unseen-tool challenge, but because the challenge set is positive-only, it cannot establish false-positive performance. This proposal preserves that distinction by reporting only recall and verdict coverage on HKD.

Jerabek et al. (2024) recreated seven DoH detectors across several data and longevity scenarios and documented transferability and usability concerns. Their comparative evidence directly supports the use of grouped partitions, explicit strata, and a locked challenge rather than a single random held-out table. Yusof et al. (2022) further showed the value of inspecting realistic IDS dataset distributions, reinforcing the planned raw-PCAP and label audit before any model is fit.

2.2.5 Evaluation Discipline: Leakage, Calibration, and Safety

Kapoor and Narayanan (2023) describe how leakage can invalidate machine-learning conclusions when information from the evaluation target enters training or model selection. Varma and Simon (2006) likewise show that using the same cross-validation process for tuning and estimating error can be biased. In packet datasets, dependence can arise from the same capture, replay, tool session, or address-controlled scenario. The planned role-separated capture groups are therefore a core validity control, not merely a software convenience.

Calibration is necessary when a predicted probability determines an action. Guo et al. (2017) distinguish discrimination from calibration and show that a model can classify well while producing unreliable probabilities. The study will fit the calibrator only after model and packet-budget selection, use a separate set to choose the action threshold, and leave safety confirmation and final testing untouched. This separation prevents threshold tuning from consuming the same benign observations used to certify the false-block bound.

Because malicious traffic may be rare, accuracy alone is not an adequate summary. Saito and Rehmsmeier (2015) recommend precision-recall analysis for imbalanced tasks, while Chicco and Jurman (2020) explain the balanced information captured by Matthews correlation coefficient. The proposal will report PR-AUC, precision, recall, MCC, confusion matrices, Brier score, calibration plots, verdict coverage, and stratified results. For blocking risk, the exact binomial method of Clopper and Pearson (1934) gives a conservative confidence interval. With zero false blocks, at least 36,887 independent eligible prefixes are needed for a two-sided 95% upper bound at or below 0.0001; this requirement prevents a small clean sample from being presented as proof of one-error-per-10,000 safety.

Random forests provide a bounded nonlinear comparison to logistic regression (Breiman, 2001), while logistic regression offers a transparent baseline. SHAP provides a consistent additive feature-attribution framework (Lundberg & Lee, 2017), but attribution must remain separate from action timing and should not be treated as causation. CRISP-DM supplies a structured lifecycle from business understanding through deployment (Chapman et al., 2000), and NIST AI RMF emphasizes validity, reliability, transparency, and risk management (National Institute of Standards and Technology, 2023). Together, these sources support a stage-gated development and evaluation process.

Source	Primary focus	Use in this proposal
Hoffman & McManus (2018)	DoH protocol standard	Defines the encrypted transport and HTTP requirements.
Thomson & Benfield (2022)	HTTP/2 standard	Supports the connection-level, non-per-query inference unit.
Bishop (2022)	HTTP/3 standard	Justifies delimiting QUIC/HTTP/3 from the TCP prototype.
Böttger et al. (2019)	DoH ecosystem and cost	Establishes practical adoption and performance context.
Siby et al. (2020)	Encrypted-DNS traffic analysis	Shows metadata signal and environment-dependent performance.
Scarfone & Mell (2007)	IDPS guidance	Frames detection, prevention, deployment, and operational impact.
CIC (2020)	CIRA-CIC-DoHBrw-2020	Primary raw-PCAP corpus and documented traffic factors.
MontazeriShatoori et al. (2020)	Time-series DoH tunnel detection	Foundational focal study and dataset task.
Vekshin et al. (2020)	ML-based DoH recognition	Demonstrates classification from flow metadata.
Singh & Roy (2022)	Ensemble malicious-DoH detection	Shows strong conventional-model benchmark results.
Zebin et al. (2022)	Explainable DoH IDS	Establishes that XAI for DoH is not the missing novelty.
Mitsuhashi et al. (2022)	Persistent tool recognition	Motivates cross-tool and persistence-aware evaluation.
Yusof et al. (2022)	Dataset visualization	Supports distribution and label auditing.
Abu Al-Haija et al. (2023)	Lightweight two-stage model	Shows feature reduction and staged classification.
Casanova & Lin (2023)	ML algorithm comparison	Confirms broad benchmark feasibility.
Mitsuhashi et al. (2023)	HKD dataset	Provides unseen dnstt, tcp-over-dns, and tuns challenge traffic.
Jerabek et al. (2024)	Comparative detector analysis	Highlights transferability, usability, and longevity.
Pich et al. (2024)	Single-packet features	Shows prior attention to early evidence.
Aggarwal & Kumar (2024)	Ensemble framework	Reinforces that another accuracy-only ensemble is insufficient.

Du et al. (2026)	Real-time explainable detection	Differentiates high-speed backbone work from this edge study.
Breiman (2001)	Random forests	Supports the bounded nonlinear comparator.
Lundberg & Lee (2017)	SHAP	Supports post-action feature attribution.
Saito & Rehmsmeier (2015)	PR evaluation	Supports imbalance-aware reporting.
Chicco & Jurman (2020)	MCC interpretation	Supports a balanced binary-classification metric.
Guo et al. (2017)	Probability calibration	Separates calibrated risk estimates from discrimination.
Kapoor & Narayanan (2023)	Data leakage	Supports role separation and prohibited predictors.
Varma & Simon (2006)	Selection bias	Supports nested, untouched evaluation roles.
Clopper & Pearson (1934)	Exact binomial interval	Defines the automatic-blocking safety gate.
Chapman et al. (2000)	CRISP-DM	Structures the stage-gated development lifecycle.
NIST (2023)	AI RMF	Frames validity, transparency, and risk controls.
Philippine EO 58 (2024)	National policy	Establishes bounded national relevance.

Table 2. Evidence map of the 31 principal sources synthesized in the review.

2.3 Conceptual Framework

The study adopts an Input–Process–Output framework. Inputs are the audited packet captures, fixed system and ruleset configuration, hardware manifest, approved safety criteria, and registered load scenarios. The process converts eligible connection prefixes into leakage-resistant features, selects and calibrates a bounded model using separated groups, applies the frozen policy during contained replay, and evaluates both detection and operational behavior. Outputs are alerts or qualified temporary blocks, asynchronous explanation records, statistical performance evidence, safety evidence, and resource-cost measurements. Feedback is limited to development partitions; final-test results do not return to model or threshold selection.

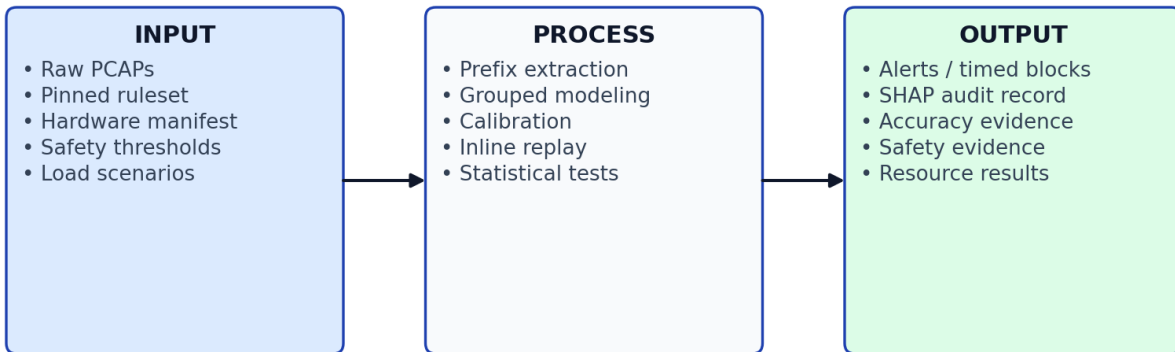


Figure 1. Input–Process–Output conceptual framework of the proposed study.

The independent experimental factor is the gateway mode: routing-only, Suricata-only, ML-only, or hybrid. Offered load is a controlled factor. Dependent variables include detection and calibration metrics, false blocks, verdict coverage, exposure before verdict, latency, throughput, loss, CPU use, and memory use. The hardware, packet corpus, connection groups, replay timing, run duration, configuration hashes, and load order policy are controlled or recorded. Approval status and the safety gate moderate whether an ML alert may become an automatic block.

2.4 Synthesis

The literature establishes four points. First, DoH provides legitimate privacy and cannot be treated as malicious merely because it is encrypted. Second, connection metadata can support classification, but environmental changes

and dataset dependence limit generalization. Third, explainable and real-time DoH detectors already exist, so neither XAI nor speed alone supplies a defensible novelty claim. Fourth, moving from an offline score to an inline block creates a separate safety and availability problem.

The proposed study responds by narrowing its claim. It will not attempt to decode a query or create a new algorithm. It will determine whether a frozen, identity-free early connection-prefix policy can add measurable value to a conventional edge stack while meeting a strict false-block gate and preserving rollback and fail-open behavior. The combination of group-separated model roles, an unseen-tool positive challenge, a pinned Suricata comparator, and measured operational cost converts the gap from a broad integration claim into a testable network-administration question.

SECTION III — METHODOLOGY

3.1 Research Design

The study will use an applied, developmental, and quantitative experimental design. It is applied because it evaluates a network-security control for a practical gateway setting; developmental because it constructs and integrates the routing, inspection, learned monitoring, explanation, and reversible-response components; and experimental because gateway mode and offered load will be controlled while outcomes are measured. The unit of offline classification is an eligible TCP connection prefix. The unit of operational replication is one ten-minute mode-load run. These units will not be mixed when calculating uncertainty.

The design has two linked experiments. The first is an offline model-development and frozen-test experiment using group-disjoint capture partitions. It will select the packet budget and model, fit a separate calibrator, tune a separate action threshold, confirm benign blocking risk, and finally evaluate untouched groups. The second is an inline contained-replay experiment comparing four gateway modes under the same registered traffic scenarios. Automatic blocking is a conditional treatment: the code path and rollback tests can be verified, but live automatic action will remain disabled unless the adviser approves the eight-week route and the safety gate passes.

Variables and Operational Measures

Class	Operational definition
Independent variables	Gateway mode (A–D); offered load (10, 50, 100 Mbps or frozen lower grid); tunnel/benign stratum; tool; resolver; capture group.
Detection outcomes	PR-AUC, precision, recall, MCC, confusion matrix, challenge recall, verdict coverage, and Suricata alerts/drops.
Calibration and safety	Brier score, calibration curve, false blocks per 10,000, exact 95% confidence intervals, action coverage, expiration success, rollback success.
Timing and exposure	Packets, bytes, and elapsed time to verdict; p50/p95/p99 forwarding latency; rule-installation latency; pre-verdict payload exposure.
Operational outcomes	Goodput/throughput, packet loss, gateway CPU utilization, memory use, queue drops, process health, and bypass events.
Controlled records	Hardware/software versions, file and configuration hashes, run duration, replay timing, run order, VM allocation, NIC/driver details, and ruleset snapshot.

Table 3. Experimental variables and measures.

3.2 System Development Methodology

Stage-Gated CRISP-DM

CRISP-DM will organize the eight-week work into business understanding, data understanding, data preparation, modeling, evaluation, and bounded deployment (Chapman et al., 2000). The stages will be treated as gates rather than as permission to revise the model after final testing. NIST AI RMF considerations—validity, reliability,

transparency, accountability, privacy, and security—will be recorded in the artifact checklist (National Institute of Standards and Technology, 2023).

Week	Gate	Deliverable
1	Approval and environment gate	Adviser approval recorded; hardware manifest completed; isolation, routing-only preflight, corpus access, hashes, and short PCAP extraction spike verified.
2	Data understanding and audit	Label reconciliation, duplicate policy, capture groups, eligibility attrition, feature schema, and partition manifest frozen.
3	Feature pipeline	First- N extraction for 20/40/60 segments, unit tests, leakage audit, and descriptive distribution review completed.
4	Model and N selection	Logistic regression and bounded random forest compared only on fitting/selection groups; one model and N frozen.
5	Calibration and threshold	Calibrator fit, action threshold selected, explanation pipeline integrated, and no further model search permitted.
6	Safety and system tests	Benign-risk confirmation, allowlist, timeout, rollback, process-stop, NFQUEUE bypass, and logging tests completed.
7	Operational experiment	Four modes $\times$ three loads $\times$ three repetitions conducted using registered order and monitoring.
8	Final analysis and documentation	Untouched final test opened, bootstrap intervals computed, limitations recorded, evidence bundle and manuscript finalized.

Table 4. Proposed eight-week stage-gated schedule, subject to adviser approval.

If the eight-week IDPS route is not approved, the project will not present the five-week option as an equal active design. Instead, it will use a documented contingency: retain data audit, feature extraction, grouped offline modeling, calibration, and final IDS evaluation; remove the HKD extension if time requires; and keep every ML response alert-only. This protects the core research validity without implying that prevention was evaluated when it was not authorized.

Sampling and Data Sources

Dataset selection is purposive because the research question requires labeled TCP-based DoH, benign DoH, ordinary HTTPS, raw packets, and known capture conditions. It is not a survey sample of all Internet traffic. Before modeling, every source file will be recorded by filename, byte size, cryptographic hash, readability, label source, capture group, and permitted role. Counts in publications or combined CSV releases will be reported as nominal context until the local raw-PCAP extractor produces an auditable eligible-prefix ledger.

Corpus	Observed classes or factors	Permitted role
CIRA primary	Benign DoH; ordinary HTTPS; dns2tcp, DNSCat2, and Iodine tunnel traffic across documented browsers and resolvers.	Fit, N /model selection, calibration, threshold tuning, safety confirmation, and final test—only through distinct capture groups.
Prospective benign-DoH safety corpus	Controlled non-personal browsing/resolution grouped by client, resolver, day, and session, with natural connection reuse.	May supplement the fixed-threshold safety set only. It may not fit, select, calibrate, or tune.
HKD locked challenge	Positive-only PCAPs generated with dnstt, tcp-over-dns, and tuns.	Recall and verdict coverage only after the model, N , calibrator, and threshold are frozen.

Table 5. Corpus roles and restrictions.

All prefixes from the same source capture or registered scenario will remain in one partition. Roles will be assigned before modeling: model fitting, model/packet-budget selection, calibration fitting, action-threshold tuning, safety confirmation, and final testing. No random row split, replay copy, or CSV augmentation will create an independent observation. If a stratum lacks enough capture groups for a role, roles will be combined only on the development side and the limitation will be disclosed; the final test will remain untouched.

Data Collection Instruments

- A corpus manifest will record source, license/access location, hash, capture group, class, tool, resolver, eligibility, and partition role.
- The prefix extractor will produce a schema-validated feature table and an attrition ledger showing why each connection was included, excluded, or marked as no-verdict.
- A run manifest will record mode, load, repetition, start/end times, configuration hashes, random seed where applicable, host health, and anomalies.
- System logs will include Suricata EVE records, nftables action and expiry records, ML scores and verdict timestamps, process-health events, and resource samples.
- Packet and traffic tools will record offered rate, received rate, loss, and latency without modifying the frozen ML feature pipeline.

3.3 Requirement Specification and Analysis

Functional Requirements

ID	Requirement
FR-01	Forward IPv4 TCP traffic between two isolated test segments without NAT or a default route to campus/public networks.
FR-02	Check explicit allowlists and bidirectional timeout block sets before established-connection acceptance.
FR-03	Queue remaining forwarded traffic to the pinned Suricata NFQUEUE path with bypass configured and logged.
FR-04	Track bidirectional TCP connections and count only unique payload-bearing segments after the handshake.
FR-05	Compute only the registered timing, direction, size, burst, and ratio features available at the frozen N .
FR-06	Apply the frozen model and calibrator; label short connections as no ML verdict.
FR-07	Write a temporary exact-five-tuple block only when route approval, threshold, allowlist, and safety status permit.
FR-08	Expire every ML block automatically and support manual rollback without restarting the gateway.
FR-09	Generate SHAP attribution asynchronously from the stored feature vector after the action decision.
FR-10	Export immutable run, configuration, decision, performance, and error records for analysis.

Table 6. Functional requirements.

Nonfunctional and Safety Requirements

ID	Requirement
NFR-01 Reproducibility	Pin and hash code, model, feature schema, ruleset, configuration, and data manifests.
NFR-02 Privacy	Do not decrypt payloads or use identity fields, SNI, certificates, or TLS fingerprints as predictors.
NFR-03 Availability	Forward safely on monitor failure; log NFQUEUE bypass; stop creating new ML blocks if controller health fails.
NFR-04 Performance	Routing-only preflight must sustain the registered top load without source/sink saturation or host-side loss.
NFR-05 Safety	Automatic ML blocking requires adviser approval and passage of both benign-stratum exact confidence gates.
NFR-06 Auditability	Every decision must link to N , features, model/calibrator versions, score, threshold, action, and expiry.
NFR-07 Containment	Replay network must have no NAT or default route to campus or public Internet; management is separate or local.

Table 7. Nonfunctional and safety requirements.

Data-Flow Models

The context model identifies the replay host, sink host, corpora, researchers, and the proposed gateway as separate entities. Test traffic crosses the gateway; corpora and frozen configurations enter the experimental process; and results leave as an evidence package. No production user or public service is part of the data flow.

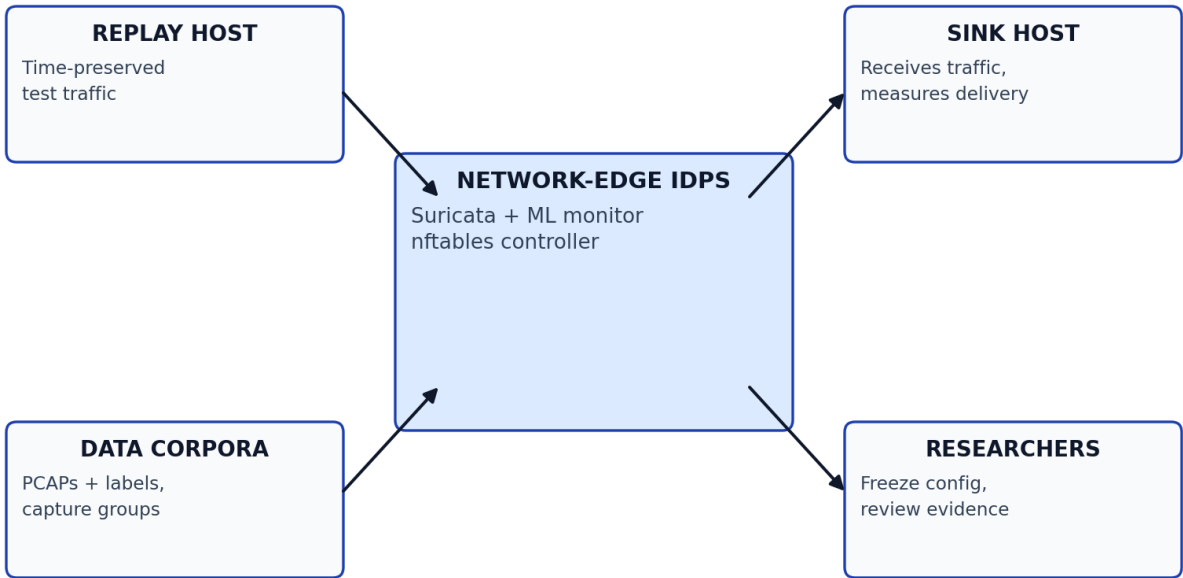


Figure 2. Level-0 system context and principal data flows.

The Level-1 model separates ingestion, feature extraction, scoring, control, and evidence recording. This separation is important because labels belong to the offline corpus process and must never enter the live feature vector. It also keeps explanation generation and analysis out of the forwarding-critical path.

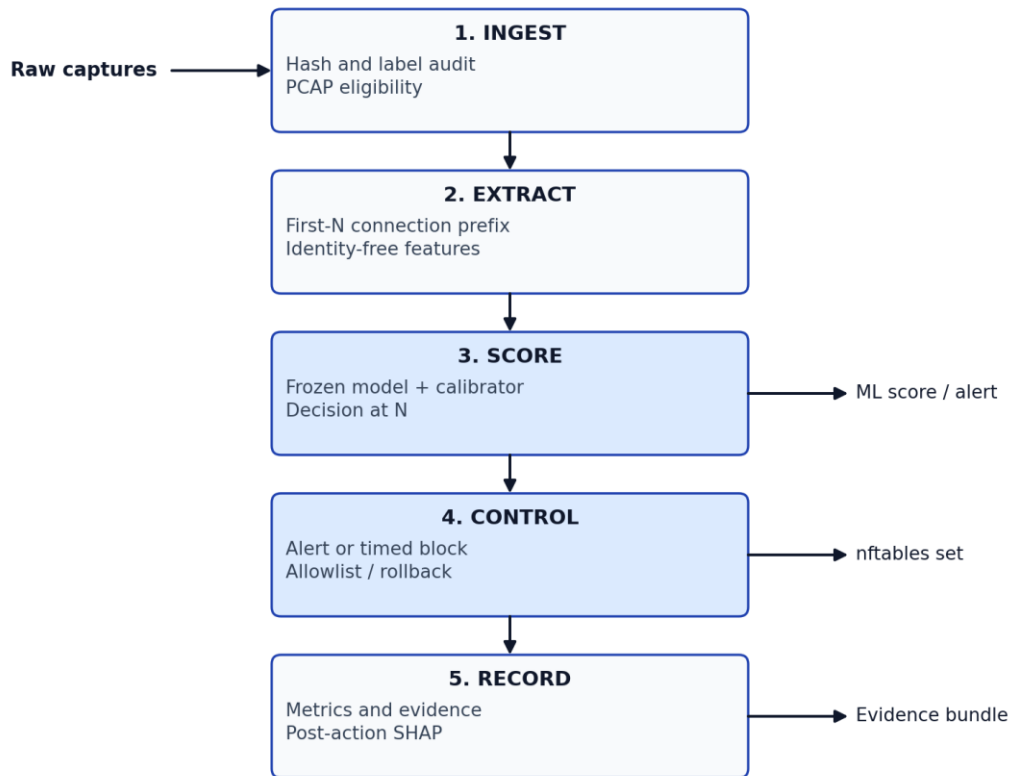


Figure 3. Level-1 data-flow model for the experimental pipeline.

Logical Data Model

The experiment does not require a production relational database, but its records will follow a normalized logical model so that decisions can be traced to capture groups without exposing predictor identities. The capture-group entity owns the data-role assignment; each eligible connection prefix has one frozen feature vector and may have one decision record for a particular model version. Run and configuration identifiers will be stored as foreign keys or manifest references in the implementation.

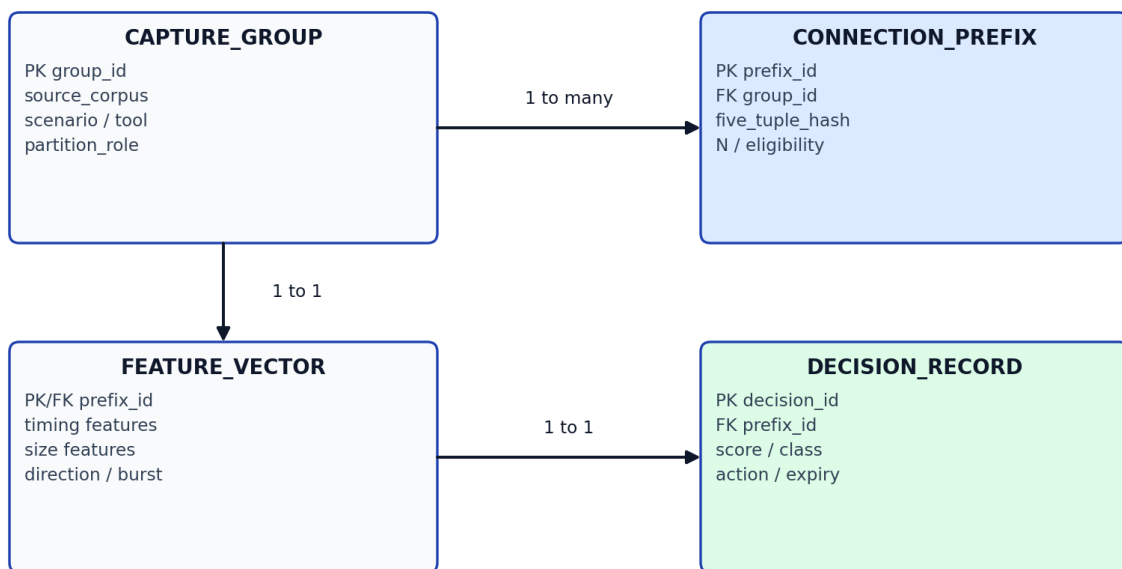


Figure 4. Core logical entities and relationships for experiment records.

Decision Flow

A connection receives an ML score only if it reaches the frozen packet budget. A score below threshold is logged and allowed. A score at or above threshold produces an alert. It becomes a block only when automatic action is enabled, the connection is not allowlisted, the safety gate passed, and the controller is healthy. This ordered check prevents the model alone from exercising unrestricted firewall authority.

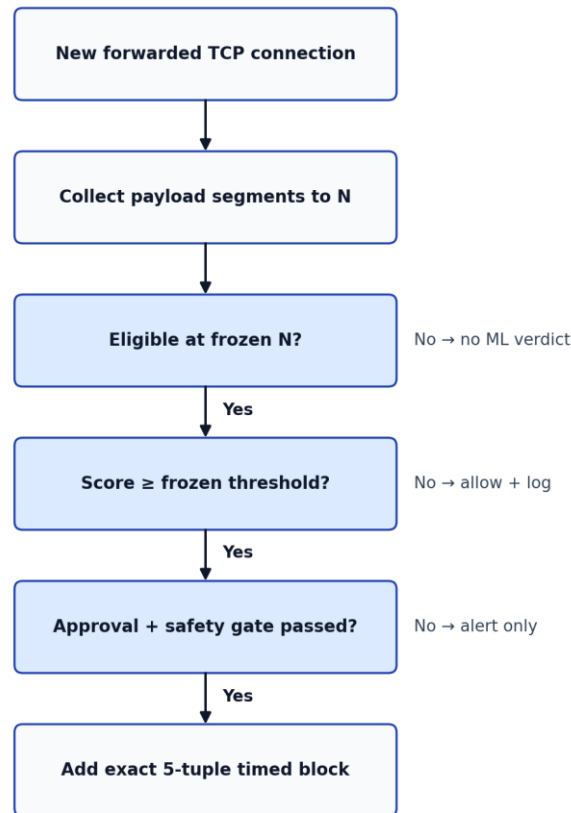


Figure 5. Early-verdict and conditional reversible-blocking flow.

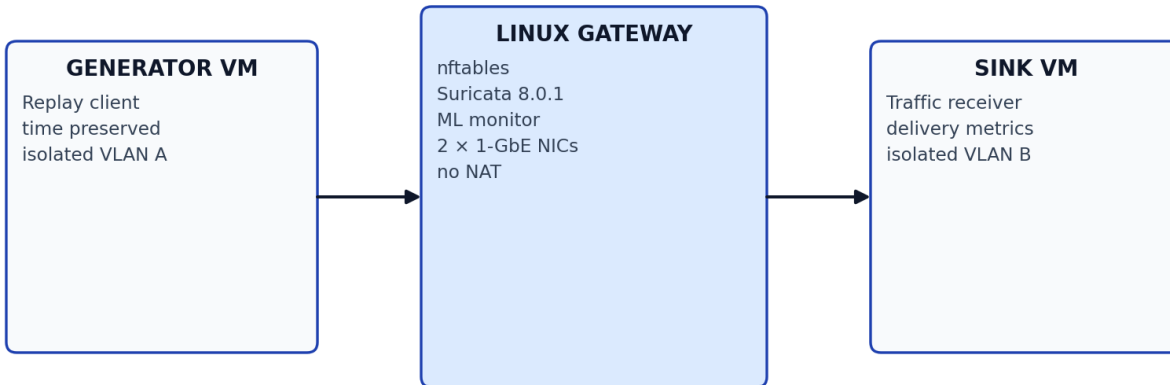
3.4 System Design Specification

Logical Specification

The gateway will use Linux IP forwarding and nftables. An early forward-chain stage will check explicit allowlists and two timeout sets representing both directions of a blocked IPv4 five-tuple. Traffic not decided by those rules will enter a dedicated NFQUEUE chain for Suricata 8.0.1. Suricata will inspect traffic using a pinned ET Open ruleset snapshot and selected drop actions. A separate passive monitor will observe the same forwarded connections, compute the frozen prefix features, produce a calibrated score, and request only expiring set updates through a least-privilege controller. The controller will reject malformed, non-TCP, out-of-scope, allowlisted, or non-expiring requests.

SHAP attribution will run asynchronously after the score and action timestamps are committed. The gateway will not wait for an explanation before forwarding or blocking. Logs will be written locally during a run and exported after completion. If the ML monitor or controller stops, no new ML block can be created and existing entries expire. If Suricata is absent, the configured NFQUEUE bypass permits forwarding and records a loss-of-inspection event.

Physical Specification



Management: local console or separate path only

No default route, NAT, campus access, or public-Internet access during replay

Figure 6. Contained two-computer routed testbed.

Component	Minimum acceptance specification
Linux gateway	At least 4 logical CPU threads; 8 GB RAM; 40 GB free SSD; two isolated 1-GbE interfaces.
Replay/sink computer	At least 4 physical/8 logical CPU threads; hardware virtualization; 16 GB RAM; 250 GB free SSD or dedicated external SSD; two 1-GbE interfaces or one 802.1Q-capable trunk.
Generator VM	At least 2 vCPU and 4 GB RAM; isolated source segment; time-preserved replay and load orchestration.
Sink VM	At least 2 vCPU and 4 GB RAM; isolated destination segment; traffic receipt and performance measurement.
Switching	Direct links or a managed switch with isolated VLANs; no shared production broadcast domain.
Management	Local console or physically/logically separate management path that does not provide replay traffic a public route.

Table 8. Minimum testbed acceptance specifications.

During Week 1, the team will replace minimums with exact make/model, CPU, RAM, storage medium, NIC chipset, driver, operating-system kernel, VM allocation, and switch/VLAN values. A proposal approval copy may state minimums, but experimental results will not retain 'TBD' values. If routing-only preflight cannot sustain the intended 100-Mbps load without source/sink saturation or host-side loss, the highest load will be lowered before any comparative run and the revised grid will be frozen.

Software and Configuration Stack

Layer	Planned specification
Operating system	Stable 64-bit Linux distribution and kernel recorded in the Week-1 manifest.
Firewall/routing	nftables and Linux IP forwarding; explicit allowlists; timeout sets; counters; no NAT.
Rule-based IPS	Suricata 8.0.1 via NFQUEUE with queue bypass; ET Open snapshot acquired through suricata-update.
Packet processing	Python-based or compiled extractor using a pinned packet library; exact versions and hashes recorded.
Machine learning	Pinned Python, NumPy, pandas, scikit-learn, joblib, and SHAP versions.
Measurement	Packet capture, traffic generation/replay, latency/goodput counters, and operating-system CPU/memory monitoring tools.
Artifacts	Read-only corpus manifest, partition file, feature schema, trained model, calibrator, threshold file, ruleset inventory, and run manifests.

Table 9. Planned technology stack; exact build versions will be frozen before experimentation.

Pinned Suricata Baseline

The Suricata comparator will be reproducible and label-independent. The experiment record will include Suricata 8.0.1, ruleset retrieval date, source URL, archive or manifest hash, enabled categories, enabled and disabled SIDs, HOME_NET and EXTERNAL_NET values, NFQUEUE number and bypass option, exceptions, and the final action map. The selected prevention rules will be converted to drop through a frozen modify configuration before any comparative run. Rules will not be added or removed after viewing final labels or ML failures.

Feature and Model Specification

Class	Specification
Permitted	Per-direction counts and bytes; payload-length minimum/maximum/mean/standard deviation; inter-arrival mean/standard deviation; direction changes; directional burst counts/sizes/durations; up/down ratios with defined zero handling.
Descriptive only	ALPN, when available, may describe the corpus but will not be a predictor.
Prohibited	IP addresses, ports, MAC addresses, browser/tool/resolver identity, SNI, certificate fields, TLS fingerprint, absolute time, filename, capture ID, final flow duration/totals, and every packet after N .
Packet counting	Unique payload-bearing TCP segments after handshake; pure ACKs and retransmissions do not advance N .
Short flows	Connections ending before N are labeled no ML verdict and included in coverage reporting.

Table 10. Registered feature boundary.

A directional burst is a same-direction run terminated by a direction change or a pre-registered inactivity gap. It is an observable traffic construct and will not be described as a DNS request or response. The same versioned extractor will process offline and live traffic, and known-PCAP parity tests must reproduce the identical feature vector before deployment.

Candidate budgets are $N = 20, 40$, and 60 . Development groups will first produce a viability report covering verdict coverage, group distributions, medians and interquartile ranges, and performance against a prevalence-matched dummy classifier. If neither candidate model exceeds the dummy on both grouped-selection PR-AUC and MCC, inline deployment will stop and the negative result will be retained. Otherwise, a class-weighted logistic regression and one bounded random forest will be compared on a small pre-declared grid. One N /model pair will be selected by grouped PR-AUC, with MCC, p95 inference time, model simplicity, and then smaller N used as ordered tie-breakers.

After model selection, a separate calibration set will choose between Platt scaling and isotonic regression by Brier score and log loss. Alert and candidate-block thresholds will then be frozen on separate threshold-tuning groups before the safety-confirmation set is opened. No XGBoost, LightGBM, deep neural network, post-safety threshold repair, or post-final-test feature search will be introduced.

3.5 Implementation

Implementation Sequence

1. Prepare the isolated testbed, verify addressing and routes, disable NAT/default routes, and record the physical/software manifest.
2. Acquire permitted corpora; calculate hashes; verify PCAP readability; reconcile published and extracted labels; define capture groups and role assignments.
3. Implement and unit-test connection reconstruction, retransmission handling, payload-bearing segment counting, first- N cutoff, feature calculations, and no-verdict behavior.
4. Run the development-only model pipeline, select one N /model, fit one calibrator, choose one threshold, and export signed or hashed frozen artifacts.

5. Configure nftables allowlists, directional timeout sets, rollback command, controller input validation, process-health checks, and audit logging.
6. Install and pin Suricata 8.0.1 and the ET Open snapshot; freeze categories, SIDs, actions, HOME_NET, EXTERNAL_NET, NFQUEUE, and exceptions.
7. Integrate the passive monitor, calibrated scorer, conditional controller, and asynchronous SHAP worker without placing training or explanation code in the forwarding path.
8. Execute component, integration, safety, and routing-only preflight tests; correct implementation defects before opening safety-confirmation or final-test groups.
9. Conduct the registered operational runs, open the untouched final groups, calculate pre-specified metrics and intervals, and archive the evidence bundle.

Team Responsibility and Review

Each major artifact will have a primary owner and a reviewer other than the owner. Responsibilities will be assigned across testbed/networking, data extraction, modeling/calibration, enforcement/safety, and analysis/documentation. Configuration changes require a recorded reason, reviewer confirmation, and a new hash. Team members may rotate operational duties, but no person will approve their own safety-gate calculation or final configuration manifest.

Configuration and Change Control

- A freeze manifest will record Git commit or source hash, environment lock file, model/calibrator checksums, threshold, N , feature order, and random seeds.
- Every Suricata and nftables configuration used in a recorded run will be copied to the run directory and hashed.
- Defects discovered after freezing will be classified as implementation errors or analytical changes. An implementation fix requires rerunning all affected modes; an analytical change requires reopening development and invalidates any previously viewed final result.
- Raw corpora remain read-only. Derived tables are reproducible from manifests and scripts; manual spreadsheet edits are not accepted as authoritative data transformation.

3.6 Testing and Evaluation

Component and Integration Testing

Test	Acceptance condition
Packet accounting	Synthetic traces with ACKs, retransmissions, reordering, short connections, and bidirectional payloads produce expected N and features.
Leakage audit	Schema test rejects every prohibited predictor and verifies no post- N timestamp, size, or total enters a feature.
Model artifact	Known vectors reproduce stored scores and probabilities within a registered numeric tolerance.
Allowlist	Allowlisted management/test connections cannot be added to an ML block set.
Exact block	Only the registered bidirectional five-tuple is interrupted; unrelated concurrent connections remain forwarded.
Timeout	Block entry expires within registered tolerance and traffic can resume without a gateway restart.
Rollback	Manual rollback removes active ML entries and records actor, time, reason, and affected key.
Monitor stop	Stopping monitor/controller prevents new ML blocks; existing timed entries expire.
Suricata stop	NFQUEUE bypass preserves forwarding and logs inspection loss.
Explanation isolation	Disabling or delaying SHAP generation does not change verdict or enforcement latency.

Table 11. Component and integration acceptance tests.

Offline Model Evaluation

The frozen final test will report the confusion matrix, PR-AUC, precision, recall, MCC, Brier score, calibration plot, and verdict coverage with denominators. Results will be disaggregated by benign ordinary HTTPS, benign DoH, tunnel tool, resolver, and capture group where counts allow. Discrimination intervals will use 2,000 capture-group bootstrap resamples with percentile 95% intervals rather than independent row resampling; the random seed will be registered before the final test is opened.

On the HKD challenge, the study will report only the number of eligible positive connections, number receiving a verdict, recall among eligible verdicts, and overall detected proportion. It will not report precision, specificity, false-positive rate, or calibration because no benign class is present. Failure on an unseen tool will be interpreted as a coverage limitation rather than repaired through challenge-set training.

Automatic-Blocking Safety Gate

Let x be the number of false blocks among n independent eligible benign connection prefixes in a safety-confirmation stratum. The policy passes only if the upper end of the two-sided 95% Clopper–Pearson interval for the binomial probability is at most 0.0001. The test is applied separately to ordinary HTTPS and benign DoH. When $x = 0$, this requires at least 36,887 eligible independent prefixes in each stratum. Replays, forced reconnects, duplicated prefixes, or augmented rows cannot increase n . If either stratum fails because of an error or insufficient evidence, automatic ML blocking remains disabled and the system is reported as alert-only.

Operational Experiment

Mode	Configuration	Purpose
A	Routing only	Linux forwarding and measurement; no Suricata inspection and no ML action.
B	Suricata IPS	Pinned Suricata rules and drop actions; ML monitor may be disabled or observation-only but cannot act.
C	ML policy	Frozen ML scoring and qualified action; Suricata enforcement disabled.
D	Hybrid	Pinned Suricata IPS plus frozen ML policy under the same routing and measurement controls.

Table 12. Registered gateway comparison modes.

The intended grid is 10, 50, and 100 Mbps, with three repetitions per mode-load combination and ten minutes per run: 4 modes $\times$ 3 loads $\times$ 3 repetitions = 36 runs and six recorded hours. Load will be increased through concurrent time-preserved connections and address rewriting inside the isolated ranges. No global replay speed-up will be used. Run order will be randomized or counterbalanced within repetitions and recorded. Between runs, temporary sets, queues, logs, and host utilization will be returned to a known baseline. Warm-up intervals, measured intervals, and cool-down procedures will be identical across modes.

Performance and Statistical Analysis

Dimension	Analysis
Detection benefit	Additional malicious prefixes detected or blocked by C and D relative to B; overlap of Suricata and ML detections; final confusion and challenge results.
Availability cost	False blocks, connection interruptions, goodput change, packet loss, p50/p95/p99 forwarding latency, rule-installation latency, queue drops, and expiry/rollback failures.
Resource cost	Median and tail CPU and memory use by gateway and relevant processes; monitoring cadence fixed across modes.
Early action	Median and percentile packets, bytes, and milliseconds to verdict; short-flow no-verdict rate; payload exposure before action.
Uncertainty	Exact binomial intervals for false-block proportions; capture-group bootstrap intervals for classification metrics; run-level summaries for operational repetitions.
Decision rule	No single accuracy target determines success. The study will report the benefit–risk profile and enforce the pre-registered blocking gate.

Table 13. Evaluation dimensions and analytical treatment.

Threats to Validity

- **Internal validity:** Capture dependence, label errors, feature leakage, tuning on the test set, replay timing changes, and uncontrolled configuration changes will be addressed through group roles, audits, frozen artifacts, time-preserved concurrency, and run manifests.
- **Construct validity:** The model detects connection-prefix behavior associated with corpus labels. It does not observe decrypted DNS content or prove that a particular query is malicious. Metrics and terminology will retain this boundary.
- **External validity:** Synthetic tools, historical software, one hardware class, TCP-only DoH, and a laboratory load grid limit generalization. The HKD challenge adds unseen tools but does not supply a benign population.
- **Conclusion validity:** Replayed flows and packet rows are not independent. Group bootstrap and run-level operational summaries will be used, denominators will be shown, and null or mixed findings will be reported.

Ethical, Legal, and Safety Controls

The study uses no human participants and will not inspect production user traffic. Public research datasets will be handled according to their access terms. Controlled prospective benign traffic will be generated by the research team without personal accounts or private content. Replay networks will be isolated, with no NAT or default route to campus or the Internet. Malicious labels will come from controlled dataset scenarios; no live malware will be executed. Results will minimize identifiers, separate labels from predictors, and retain only artifacts necessary for verification. Any unexpected route, external packet, or containment failure will stop the experiment and invalidate the affected run.

Expected Research Output

The final output will be a reproducible evidence package containing the hardware and software manifest, corpus and hash inventory, group-partition file, feature schema, leakage audit, frozen model and calibrator, threshold and safety-gate record, pinned Suricata inventory, nftables policy, functional-test results, run manifests, performance tables, calibration and attribution plots, and a bounded conclusion answering the four research questions. A positive result will require measurable incremental detection with an acceptable operational profile and, for prevention, formal passage of the blocking gate. A negative or alert-only result will still be valid if it is produced by the registered method and its limitations are reported.

REFERENCES

Journal Articles and Conference Papers

- Abu Al-Hajja, Q., Alohal, M., & Odeh, A. (2023). A lightweight double-stage scheme to identify malicious DNS over HTTPS traffic using a hybrid learning approach. *Sensors*, 23(7), 3489. <https://doi.org/10.3390/s23073489>
- Aggarwal, A., & Kumar, M. (2024). An ensemble framework for detection of DNS-over-HTTPS (DoH) traffic. *Multimedia Tools and Applications*, 83, 32945–32972. <https://doi.org/10.1007/s11042-023-16956-9>
- Böttger, T., Cuadrado, F., Antichi, G., Fernandes, E. L., Tyson, G., Castro, I., & Uhlig, S. (2019). An empirical study of the cost of DNS-over-HTTPS. In *Proceedings of the 2019 Internet Measurement Conference* (pp. 15–21). Association for Computing Machinery. <https://doi.org/10.1145/3355369.3355575>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Casanova, L. F. G., & Lin, P.-C. (2023). Malicious network traffic detection for DNS over HTTPS using machine learning algorithms. *APSIPA Transactions on Signal and Information Processing*, 12(2), 1–24. <https://doi.org/10.1561/116.00000058>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 6. <https://doi.org/10.1186/s12864-019-6413-7>
- Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4), 404–413. <https://doi.org/10.1093/biomet/26.4.404>
- Du, J., Liu, S., Zhi, C., Zhao, J., Wu, H., & Cheng, G. (2026). TunnelEye: An explainable framework for real-time detection of malicious DoH traffic in high-speed backbone networks. *IEEE Transactions on Network and Service Management*. <https://doi.org/10.1109/TNSM.2026.3682686>

- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR. <https://proceedings.mlr.press/v70/guo17a.html>
- Jerabek, K., Hynek, K., & Rysavy, O. (2024). Comparative analysis of DNS over HTTPS detectors. *Computer Networks*, 247, 110452. <https://doi.org/10.1016/j.comnet.2024.110452>
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. <https://doi.org/10.1016/j.patter.2023.100804>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).
- Mitsuhashi, R., Jin, Y., Iida, K., Shinagawa, T., & Takai, Y. (2022). Malicious DNS tunnel tool recognition using persistent DoH traffic analysis. *IEEE Transactions on Network and Service Management*. <https://doi.org/10.1109/TNSM.2022.3215681>
- MontazeriShatoori, M., Davidson, L., Kaur, G., & Habibi Lashkari, A. (2020). Detection of DoH tunnels using time-series classification of encrypted traffic. In *2020 IEEE DASC/PiCom/CBDCCom/CyberSciTech* (pp. 63–70). <https://doi.org/10.1109/DASC-PiCom-CBDCCom-CyberSciTech49142.2020.00026>
- Pich, R., Sreng, S., & Colin, J.-N. (2024). DoH tunneling traffic detection based on single packet features analysis. In *Proceedings of the 2023 12th International Conference on Networks, Communication and Computing*. <https://doi.org/10.1145/3638837.3638861>
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- Siby, S., Juarez, M., Diaz, C., Vallina-Rodriguez, N., & Troncoso, C. (2020). Encrypted DNS → privacy? A traffic analysis perspective. In *Proceedings of the Network and Distributed System Security Symposium*. <https://www.ndss-symposium.org/ndss-paper/encrypted-dns-privacy-a-traffic-analysis-perspective/>
- Singh, S. K., & Roy, P. K. (2022). Malicious traffic detection of DNS over HTTPS using ensemble machine learning. *International Journal of Computing and Digital Systems*, 11(1), 1061–1069. <https://doi.org/10.12785/ijcds/110185>
- Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, 91. <https://doi.org/10.1186/1471-2105-7-91>
- Vekshin, D., Hynek, K., & Cejka, T. (2020). DoH Insight: Detecting DNS over HTTPS by machine learning. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*. <https://doi.org/10.1145/3407023.3409192>
- Yusof, M. H. M., Almohammed, A. A., Shepelev, V., & Ahmed, O. (2022). Visualizing realistic benchmarked IDS dataset: CIRA-CIC-DoHBrw-2020. *IEEE Access*, 10, 94624–94642. <https://doi.org/10.1109/ACCESS.2022.3204690>
- Zebin, T., Rezvy, S., & Luo, Y. (2022). An explainable AI-based intrusion detection system for DNS over HTTPS (DoH) attacks. *IEEE Transactions on Information Forensics and Security*, 17, 2339–2349. <https://doi.org/10.1109/TIFS.2022.3183390>

Books, Standards, and Technical Reports

- Bishop, M. (2022). *HTTP/3* (RFC 9114). Internet Engineering Task Force. <https://www.rfc-editor.org/rfc/rfc9114.html>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS.
- Hoffman, P., & McManus, P. (2018). *DNS queries over HTTPS (DoH)* (RFC 8484). Internet Engineering Task Force. <https://www.rfc-editor.org/rfc/rfc8484.html>
- National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). <https://doi.org/10.6028/NIST.AI.100-1>
- Scarfone, K., & Mell, P. (2007). *Guide to intrusion detection and prevention systems (IDPS)* (NIST SP 800-94). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-94>
- Thomson, M., & Benfield, C. (2022). *HTTP/2* (RFC 9113). Internet Engineering Task Force. <https://www.rfc-editor.org/rfc/rfc9113.html>

Datasets, Policy Documents, and Online Media

Canadian Institute for Cybersecurity. (2020). *CIRA-CIC-DoHBrw-2020 dataset*. University of New Brunswick. <https://www.unb.ca/cic/datasets/dohbrw-2020.html>

Mitsuhashi, R., Jin, Y., Iida, K., Shinagawa, T., & Takai, Y. (2023). *DoH-Tunnel-Traffic-HKD dataset*. Hokkaido University Collection of Scholarly and Academic Papers. <https://hdl.handle.net/2115/88092>

Office of the President of the Philippines. (2024, April 4). *Executive Order No. 58: Adopting the National Cybersecurity Plan 2023–2028 and directing its implementation*. Official Gazette of the Republic of the Philippines. <https://www.officialgazette.gov.ph/2024/04/04/executive-order-no-58-s-2024/>

Open Information Security Foundation. (2026). *Setting up IPS/inline for Linux: Suricata 8.0.1 documentation*. <https://docs.suricata.io/en/suricata-8.0.1/ips/setting-up-ipsinline-for-linux.html>

Republic of the Philippines. (2012). *Republic Act No. 10173: Data Privacy Act of 2012*. Official Gazette. <https://www.officialgazette.gov.ph/2012/08/15/republic-act-no-10173/>

United Nations. (n.d.). *Goal 9: Industry, innovation and infrastructure*. <https://sdgs.un.org/goals/goal9>