Practical Advances in Artificial Intelligence and Computing
Frontier Research Publishing • https://www.frpubs.org
Vol. 5, No. 7, 2026

Network Intrusion Detection under Changing Traffic and Delayed Labels

Serovelle Wrenford[1]*, Aldric Thornevale[1], Ilyra Mosswick[1]
[1]University of Windsor, Windsor, Canada
*Corresponding author: serovellewrenford815@outlook.com

Abstract: A transparent historical benchmark experiment is used to distinguish intrusion classification from adaptive deployment claims. A fixed ridge-margin detector is fitted to 25,192 public NSL-KDD training records and evaluated on 22,544 designated test records. Training-only scaling and categorical levels are used, and the final difficulty field is excluded. Test accuracy is 0.76260, attack recall is 0.63415, and the normal false-positive rate is 0.06766. Recall is 0.44776 for 3,752 attack rows whose source labels are absent from training. Analytic prevalence calculations expose alert-workload consequences. Literature through 2026 motivates chronological replay and independent candidate promotion. These results do not estimate current-network performance.

Keywords: intrusion detection; NSL-KDD; ridge baseline; concept drift; delayed labels

1 Research Question and Empirical Scope

Network intrusion detection is a supervised learning problem only after the observation unit, labeling process, and deployment time have been specified. A row in an intrusion benchmark is not a complete account of network behavior, and a classifier score is not an operational incident. This study asks how a transparent baseline behaves when moved from a historical training subset to its designated test subset, and what that result implies for adaptation to changing traffic. It combines an actually executed ridge-margin experiment on public NSL-KDD files with a literature-based design for chronological model review.

Recent 2026 preprints investigate latent-space drift adaptation, graph-based drift mitigation, and validation before promoting an adaptive challenger [1], [2], [3]. Their shared motivation is relevant: changes in the observed traffic distribution do not by themselves prove that a newly fitted detector is better. The empirical experiment here does not reproduce those systems. It provides a fixed, auditable reference and a concrete example of the gap between pooled benchmark accuracy, recognition of unfamiliar attack labels, and the expected workload of an alert queue.

The dataset is deliberately historical. Analysis of KDD-derived data documented limitations that motivated NSL-KDD [4]. Other benchmark efforts, including UNSW-NB15, were designed under different traffic and labeling conditions [5]. Concept drift research provides a vocabulary for changes in relationships over time [6], while the base-rate problem explains why rare attacks are difficult to detect usefully [7]. Earlier criticism of closed-world evaluation remains directly relevant to deployment [8]. A 2026 manuscript using these resources should identify their age and scope rather than imply that they represent a current production network.

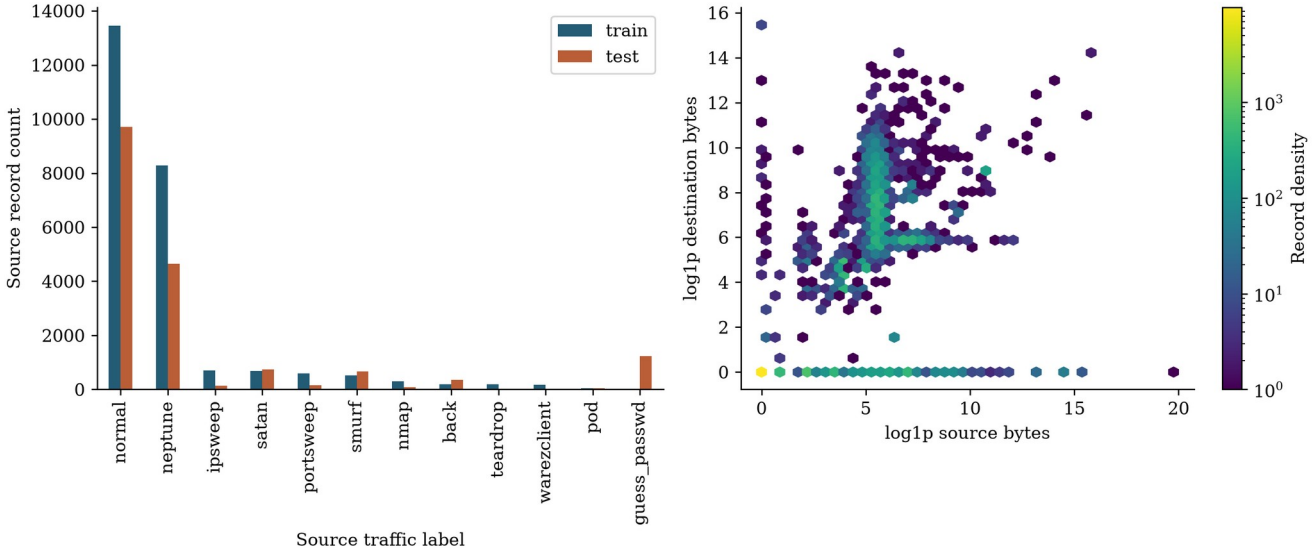


Figure 1. Counts for the 12 largest training labels in both NSL-KDD files, and training-record density over log1p byte fields. The source sample contains 25,192 training rows and 22,544 test rows; its mixture is not a measured current-network prevalence.

The public files analyzed here contain 25,192 training rows and 22,544 test rows. The source attack-label vocabulary has 22 labels in training and 38 in testing, including the normal label. The binary target is whether the original label differs from normal. No attack traffic is generated, and no live network is scanned. All calculations operate on the downloaded public benchmark records. The scripts retain predictions and source labels so that the reported results can be recomputed.

2 Dataset Audit before Fitting

The first forty-one source columns are used as candidate predictors. Three are categorical: protocol type, service, and flag. The remaining thirty-eight are parsed as numeric. The original traffic label defines the target. The final difficulty field is excluded from predictors. Including that field would give the model information about the benchmark construction rather than a deployment observation. The audit records the original column meanings and the exact transformation so that a reviewer can identify such leakage.

Categorical levels are learned from the training file only. One-hot encoding uses those levels for both subsets. A category not observed in training has no active indicator among the learned levels. That is a declared fallback, not evidence that the unfamiliar category is benign. The resulting design has 118 columns before the intercept, consisting of thirty-eight numeric predictors and categorical indicators. Training means and population standard deviations are used for standardization, with a unit scale assigned to a constant column.

Figure 1 reports actual source label counts and the joint distribution of logged byte counts in the training records. The logarithm is a display transformation using log1p, which handles zeros. It is not applied as a hidden change to the model features. The count chart is descriptive; it does not estimate attack frequencies on a real network. The distribution reflects a curated benchmark with its own sampling procedure.

The training attack fraction is 0.46614, while the test fraction is 0.56924. These are facts about the source files. Figure 2 compares empirical cumulative distributions for two numeric fields. A difference can reflect the benchmark's construction and class mixture as well as a change in conditional traffic behavior. It should

not automatically be described as chronological concept drift. The files do not provide the continuous deployment history needed for that stronger claim.

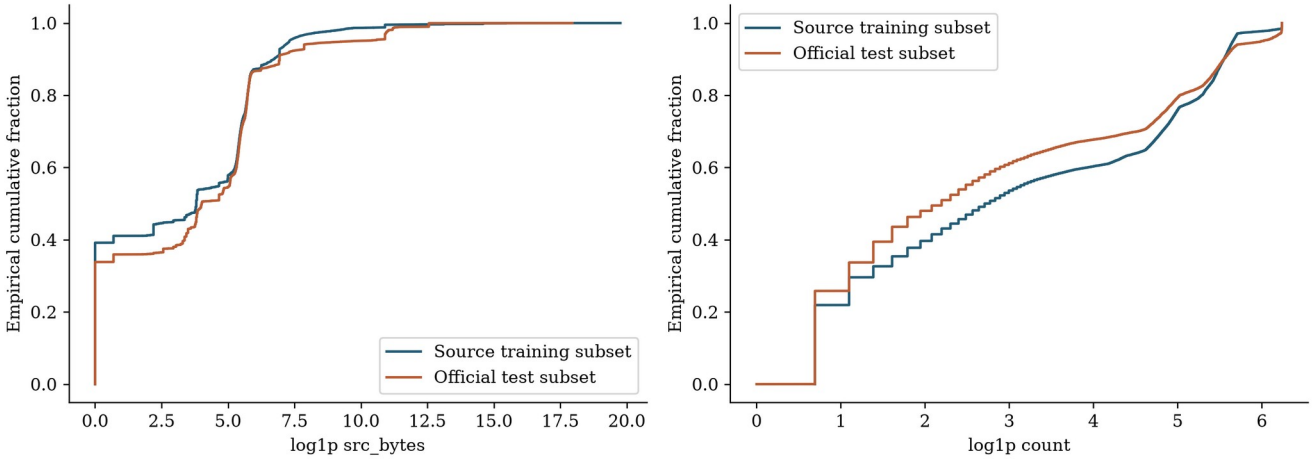


Figure 2. Empirical cumulative distributions of two actual source fields in the designated subsets. Log1p is used only for display. Differences between benchmark files are not by themselves a demonstration of chronological concept drift.

Table 1. Observation quantities and their reference requirements.

Quantity or input	Required evidence	Interpretation
Numeric fields	Thirty-eight source predictor columns	Training-only centering and scaling
Categorical fields	Protocol service and flag	Levels learned from training only
Binary target	Original normal or attack label	Does not identify individual attack families
Excluded difficulty field	Final benchmark metadata column	Excluded because it is not a traffic predictor

3 A Fixed Ridge Margin Baseline

The fitted model minimizes squared error to the binary attack indicator with an L2 penalty. Equation (1) gives the objective. The intercept is unpenalized, and the penalty is fixed at one before evaluating the test file. No test-based tuning is performed. The fitted real-valued output is a margin, not a calibrated attack probability. The classification threshold is fixed at 0.5. Clipping or passing that margin through a sigmoid would not, by itself, establish probabilistic calibration.

$$min_{b,w}\Sigma_i(y_i - b - x_i^T w)^2 + \lambda||w||^2, \lambda = 1 \quad (1)$$

The solution uses a direct regularized linear system after training-only standardization. All preprocessing, fitting, and evaluation are deterministic. This baseline is chosen for transparency and reproducibility, not to claim state-of-the-art intrusion detection. Its purpose is to expose the information conditions and error structure that a more complex model should improve under a common protocol. A useful neural comparison would use the same source fields, label exclusions, and test partition.

The test confusion matrix contains 9,054 correctly accepted normal rows, 657 normal rows flagged as attacks, 4,695 missed attack rows, and 8,138 correctly flagged attacks. Overall accuracy is 0.76260, balanced

accuracy is 0.78325, attack recall is 0.63415, and the normal false-positive rate is 0.06766. Figure 3 shows that matrix and the actual margin distributions. These results concern the designated NSL-KDD test file. They are not estimates of a modern operational network's security performance.

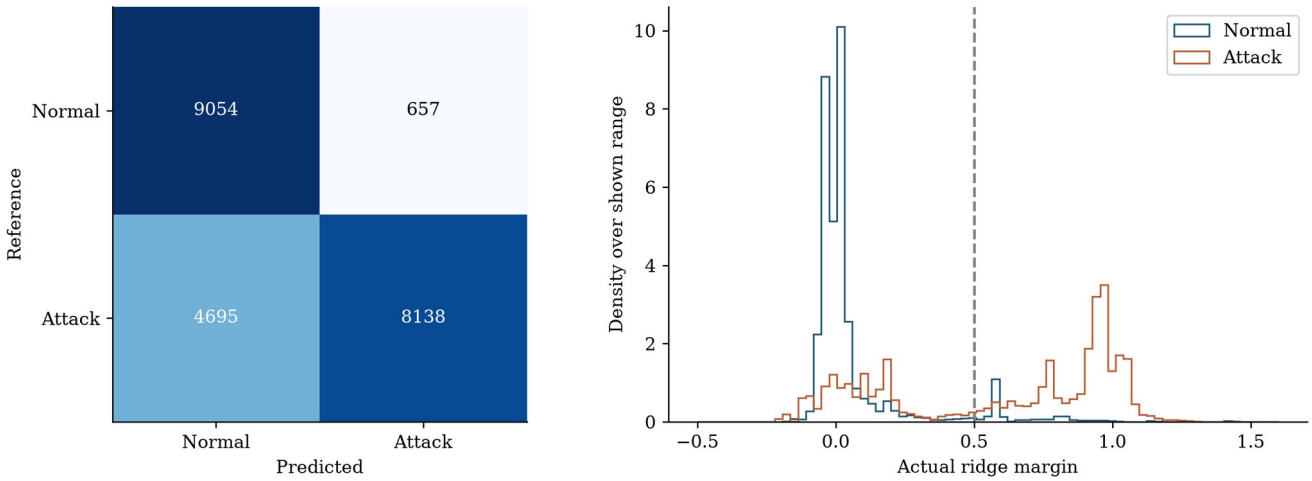


Figure 3. Actual fixed ridge-margin test confusion matrix and margin distributions. Predictor scaling and categorical levels are fitted on training records only; penalty is one and the classification threshold is 0.5. Margins are not calibrated probabilities.

The test file includes 3,752 attack rows whose specific source labels do not occur in training. Their binary attack recall is 0.44776. The model was not trained to identify individual attack families, so this subgroup evaluation concerns binary detection of rows carrying unfamiliar labels. It does not establish complete open-set attack recognition. The distinction is useful: pooled accuracy can conceal a substantial deterioration on a subgroup that is closer to the intended novelty question.

Table 2. Actually executed NSL-KDD ridge-margin results.

Quantity	Value	Evaluation scope
Training rows	25,192	Public 20 percent training file
Test rows	22,544	Designated public test file
Accuracy	0.76260	Binary normal versus attack
Attack recall	0.63415	All test attack rows
Normal false-positive rate	0.06766	All test normal rows
Absent-label attack recall	0.44776	3,752 test attack rows with labels absent in training

4 What a More Complex Model Must Demonstrate

Random forests and support vector machines provide strong conventional comparisons for tabular fields [9], [10]. Residual feature learning can be considered when a justified structured representation is available [11]. Recurrent and attention-based models are relevant when the input preserves an actual sequence [12], [13]. A table of independently sampled flow rows should not acquire a temporal interpretation merely

because its rows are fed to a sequence model. The sequence order and available history must correspond to observations the deployed system can obtain.

Autoencoder representations can support reconstruction or anomaly features [14]. Their reconstruction error is not automatically an attack probability. Focal loss can modify the fitting emphasis under class imbalance [15]. Temporal convolutional designs offer another way to summarize a valid history [16]. Each method should be assessed under the same input availability and label policy. A model given complete session outcomes or future traffic aggregates has a different information setting from one required to decide at session start.

Transfer theory clarifies why performance in one traffic collection does not guarantee performance in another [17]. Boosted trees provide a competitive tabular reference [18]. A cross-dataset study should first align field definitions and measurement processes; similarly named columns may not be equivalent. Grouped and chronological evaluation are needed when rows share hosts, sessions, or collection periods [19]. Model-comparison uncertainty should reflect those groups [20], rather than treating every correlated record as an independent operational trial.

The present baseline does not provide that full deployment evaluation. It provides a reproducible numerical anchor. A stronger model should improve relevant errors on the same fixed files before claims are broadened, then be evaluated on later independent traffic with a documented acquisition and labeling process. Better performance on the historical benchmark and better adaptation to a real stream are separate empirical questions. Both are worth studying, and neither should substitute for the other.

5 Alerts under a Different Base Rate

Calibration is a property to be measured under a declared distribution [21]. Proper scoring rules assess probability forecasts when those are actually issued [22]. The ridge margin in this experiment remains a margin. Figure 4 therefore uses sensitivity and false-positive rate in an analytic prevalence calculation rather than pretending to plot calibrated model probabilities. Equation (2) gives expected alert precision under an assumed deployment attack prevalence pi.

$$PPV(\pi) = TPR \cdot \pi / [TPR \cdot \pi + FPR \cdot (1 - \pi)] \tag{2}$$

The figure varies prevalence and several assumed false-positive rates. One curve uses the measured benchmark false-positive rate as a conditional input to the calculation. This is a scenario, not a prediction that the same rate will persist on another network. The purpose is to show the consequence if it did. A high-volume mostly normal stream can produce a large number of false alerts even when a benchmark metric appears satisfactory. Operational evaluation must therefore report exposure volume and review capacity alongside classification quality.

Equation (3) expresses an expected false-alert count for N eligible observations. It assumes the stated normal fraction and false-positive rate apply to those observations. Real alerts can be aggregated into incidents, suppressed by rules, or correlated within a session. A deployed system should measure that entire policy. The calculation is still useful for planning because it makes the denominator visible and prevents a row-level false-positive rate from being discussed as though it had no workload consequence.

$$E(falsealerts) = N(1 - \pi)FPR \tag{3}$$

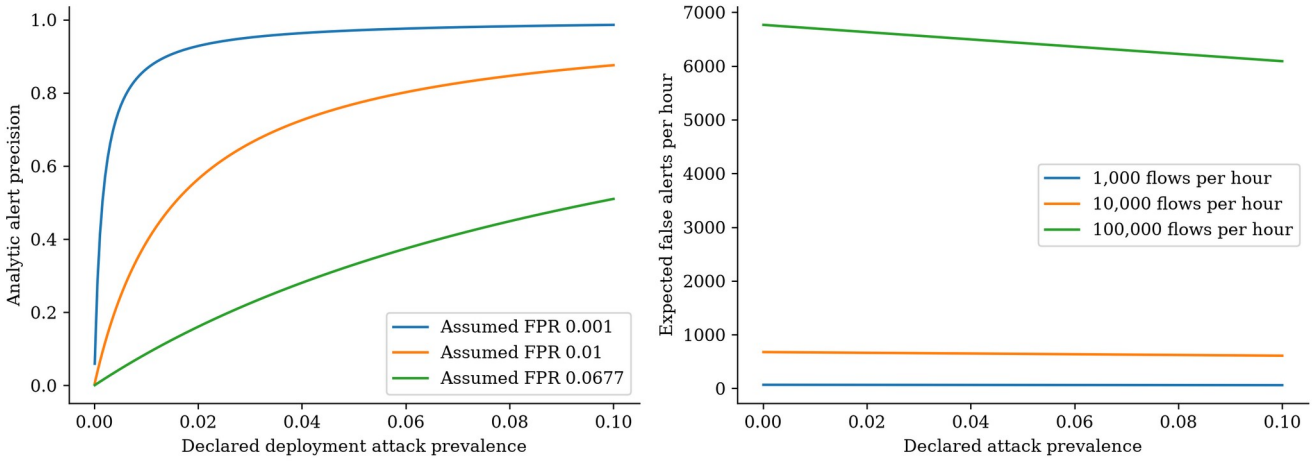


Figure 4. Analytic prevalence and workload scenarios. One assumed false-positive rate equals the measured benchmark value, but transport of that rate to another network is not claimed. Exposure rates and prevalence are declared inputs.

Ensembles can reveal variation between fitted detectors [23]. Feature attribution can help inspect a decision's dependence on particular fields [24]. Neither establishes that a flagged row is an incident. Shortcut learning can involve collection artifacts or source-specific identifiers [25]. A useful review connects the flag to the original measurement context and an independent labeling process. The operational outcome should be evaluated after aggregation and review, rather than inferred from a classifier's confidence alone.

6 Drift Detection and Model Promotion

A drift detector asks whether selected observations differ from a reference distribution. It does not directly ask whether an attack is present or whether a replacement classifier will be better. A software rollout, new service, altered logging rule, or genuine attack campaign can each change traffic summaries. The response should depend on a diagnosis and available labels. Automatically retraining on all newly observed data can teach the model to normalize a persistent failure or reinforce its own incorrect pseudo-labels.

Hyperparameter search belongs inside a development period [26]. Distribution tests, including kernel-based comparisons, can support a drift audit [27]. Their thresholds should account for repeated testing and the consequences of alarms. A stable regularized baseline offers a useful incumbent [28]. A candidate update should be compared with that incumbent on a later labeled set representative of the conditions that triggered concern. The 2026 work on conditional validation before promotion is relevant to this distinction [3].

Delayed labels complicate the comparison. At issue time, some apparently normal rows may not yet have been investigated. A deployment record should retain label availability times and later revisions. A retrospective replay should reveal which labels were actually known at each update. Otherwise, the model can receive future knowledge while the evaluation is described as online. The resulting performance would measure a more favorable information setting than the deployed system had.

Equation (4) defines a simple net utility under declared benefits and costs for correct detections, false alerts, and misses. It is a decision model, not a universal security value function. Costs can vary by asset and incident type. The equation helps expose those assumptions so that a threshold or promotion decision can be reviewed. A paper should report the underlying counts as well as any utility summary, allowing readers to apply a different reasonable cost structure.

$$U = a \cdot TP - b \cdot FP - c \cdot FN \tag{4}$$

Table 3. Failure mechanisms and records needed for interpretation.

Evidence or condition	Required comparison	Interpretation
Late discoveries and revision of benign labels	Error by label age and revision status	Label versions and availability timestamps
New services categorical levels and logging failures	Coverage and errors among accepted observations	Missingness flags and categorical vocabularies
Drift alarms without confirmed performance gains	Paired misses false alerts and workload	Candidate versions cutoffs and promotion criteria
Correlated alerts and unequal asset consequences	Utility components and review burden	Underlying contingency counts and cost assumptions
Different protocols logging and normal behavior	Coverage calibration and condition specific errors	Collection metadata feature mappings and complete predictions

7 Reproducibility and Practical Limits

The accompanying package retains the two source files, their URLs, the transformation code, source class counts, predictions, and subgroup membership. Re-running the script reconstructs the actual confusion matrix and the plotted distributions. The model excludes the final difficulty field and uses training-only scaling and categorical levels. These details make the numerical result reviewable. They do not resolve the age or sampling limitations of the benchmark, which remain part of the interpretation.

A future streaming study should add session and host grouping, measurement timestamps, label availability, incident aggregation, and review outcomes. It should evaluate a frozen incumbent alongside any adaptive candidate over the same later interval. Report missed attack types, normal false alerts, review time, and observation coverage. A detector can be useful under a constrained operating envelope without being a universal solution for changing networks. The evidence should identify that envelope directly.

8 Chronological Validation and Promotion Studies

The following streaming tests extend the historical benchmark experiment into a proposed deployment evaluation. They have not been executed on a live network in this manuscript. They separate feature provenance, delayed labels, independent sessions, causal replay, novel categories, adaptation, and incident decisions. A future study should implement them on an authorized collection with its logging and review process intact. The historical result then serves as a transparent reference rather than a substitute for that external evidence.

8.1 Feature provenance

The first requirement in feature provenance is a traceable observation. The experimental unit is a source flow record and its derived predictors, and the quantity of interest is a predictor available at the stated decision time. The main factor to audit is postevent labels and difficulty metadata. This test should begin with the original measurement record, before fitted scores or attractive examples are examined. A source value should retain its units, sampling support, and availability status. If a derived quantity changes, the study

should be able to identify the transformation responsible. Otherwise, a difference attributed to learning can originate in the measurement pipeline. The reference should be logging definitions and source column documentation, with its own uncertainty stated rather than assumed away.

The proposed comparison is to compare permitted inputs with a separately audited exclusion list. This is a prospective protocol, not an additional experiment reported as completed here. It should use the same eligible units and the same target convention on both sides. Acquisition and transformation choices should be fixed before the final evaluation. The diagnostic is leakage checks and feature availability. Examine its distribution across the actual independent units, retaining failures and unsupported observations in the denominator where the policy requires them. Report the raw counts or residuals alongside any normalized summary, because an identical percentage can represent very different evidential support in two collections.

The key interpretive limit is that a numeric column can encode benchmark construction rather than traffic. If the comparison changes both physical support and the learned representation, it cannot isolate an architectural benefit. A useful sensitivity analysis changes one documented component at a time and reports whether the conclusion persists under reasonable alternatives. Preserve feature definitions and exclusion decisions. These records connect the calculation to the source and make an independent reconstruction possible. A technically successful result would identify the conditions under which the observation remains usable, the conditions under which it fails, and the additional measurement needed to resolve those failures. It would not infer a wider application merely from a clean numerical output.

8.2 Delayed reference labels

The reference study for delayed reference labels concerns a traffic record and its later investigation. It asks whether a label supported by an independent review process can be supported consistently when the evidence is influenced by late discoveries and revision of benign labels. Begin with a written reference convention and a sample chosen independently of model confidence. The reference is investigation records carrying label availability times. Reviewers should have the information needed for that convention, but an intended independent judgment should not be anchored by the model's proposed answer. Retain disagreements and the reason for uncertainty. Consensus can be useful, yet it should not erase the fact that some observations support more than one reasonable interpretation.

To investigate this question, compare issue time labels with later adjudicated outcomes. Score error by label age and revision status, and distinguish uncertainty in the reference from disagreement with a definite label. An ambiguous case can be evaluated through admissible alternatives, an uncertainty interval, or an explicitly excluded category, provided the choice is fixed in advance. Exclusion changes the evaluated population and must remain visible. The study should also measure the effect of the reference policy on the substantive conclusion. A small average disagreement can matter if it concentrates in rare or consequential cases. Conversely, tiny discrepancies may be scientifically immaterial under the intended measurement tolerance.

Interpretation should respect the fact that an uninvestigated observation is not proven benign. The purpose of review is to establish what the observation supports, not to make every case fit a binary success story. Compare the result under reasonable reference conventions without selecting the most favorable one after seeing the model. Preserve label versions and availability timestamps, including the unresolved cases and the final adjudication history. These materials allow later researchers to distinguish a model improvement from a change in annotation practice. They also identify where a better instrument, another observation, or domain expertise would provide more value than further optimization against an uncertain label.

8.3 Independent sessions

The generalization test for independent sessions uses a complete host session or collection block as the unit that must remain intact. Its target is transfer beyond related sampled rows, while the dependence of concern is shared hosts services and repeated sessions. A random split of convenient numerical rows can conceal this dependence. The partition should therefore be created at the source-unit level before overlapping crops, windows, or augmented records are derived. Each child record should inherit its parent's assignment. The independent reference is independent group labels and collection metadata. This establishes an evaluation population that corresponds to the claim rather than to the easiest available data arrangement.

The proposed experiment is to split complete groups before fitting preprocessing or representations. Evaluate group conditional attack recall and false positives for the held-out units, showing their range as well as a pooled value. The split should also state which kinds of novelty remain shared across the boundary. Holding out one axis does not automatically hold out every other axis. If the intended use requires several changes at once, assess that combined condition explicitly and use additional comparisons to diagnose the source of any deterioration. Development decisions should be made without inspecting the locked outer reference labels.

The main limit is that row count can exaggerate independent security evidence. An uncertainty calculation should resample or otherwise model the units that are independent for the scientific question. Treating correlated children as independent trials can make a small difference appear decisive. Preserve session grouping and split assignments, including the construction of every derived record. A successful test would support a statement about the chosen transfer axis and the represented population. It would also identify whether errors concentrate in particular groups. This is more informative than a single favorable split that cannot be related to a concrete future observation setting.

8.4 Chronological replay

Timing in chronological replay should be defined for a sequence of detector issues and updates. The intended statement is a prediction using only available history, and the potential mismatch is future aggregates and hindsight labels. Record when the physical observation occurred, when the processed input became available, and when the model or decision was issued. These timestamps can differ substantially. A retrospective array can contain information unavailable at the nominal issue time even when its elements carry earlier labels. The timing reference is timestamped logging and delayed investigation records. The test should reconstruct the allowed information state rather than assuming that a correctly sorted table is necessarily causal.

The proposed comparison is to replay traffic in order with explicit feature and label cutoffs. Measure causal error and issue delay under the actual buffering and processing policy. A result should retain the exposure and issue schedule, including intervals that produce no usable output. Delayed reference information must not enter an earlier model update. If the target interval overlaps successive issues, the dependence should remain part of verification. A useful ablation changes the permitted history or waiting rule while holding other conditions fixed. This separates an improvement due to additional observation from an improvement due to a better representation of the same observation.

The interpretation is constrained because a static benchmark split is not a continuous deployment history. Preserve issue timestamps and update manifests, so that another analysis can replay the issue sequence and verify what the system could have known. Report the distribution of delays, not merely the average forward-pass time of one model call. Long tails and missed deadlines can determine usefulness for a time-sensitive

action. A defensible result identifies the time at which information becomes actionable and the conditions under which the intended interval is no longer supported. It does not infer timely operation solely from accurate retrospective localization or classification.

8.5 Unavailable fields

The incomplete-observation study for unavailable fields concerns a record with missing or unseen predictor values. It tests a declared fallback for unsupported inputs when observations are affected by new services categorical levels and logging failures. Missingness should have an explicit representation and provenance. A numeric placeholder is not automatically a physical zero, and an absent output is not automatically a negative observation. Define the eligibility rule and fallback before evaluating results. The reference is independent logging status and reference labels, which should distinguish unavailable measurement from a genuine observed absence. This separation is necessary to keep both coverage and accuracy interpretable.

The experiment should compare rejection missing indicators and training only encoding. Evaluate coverage and errors among accepted observations jointly, because a method can improve accepted accuracy simply by declining difficult inputs. Record the reasons for rejection and examine whether they are associated with the target or with a particular population. A constructed deletion test is useful for isolating an algorithmic dependency, but it should remain identified as constructed. Naturally occurring interruption may have a different mechanism and may correlate with difficult environmental or operational conditions. Both forms of evidence can be useful if their roles are kept separate.

The limit is that all zero categorical indicators do not establish benign traffic. A robust system should make its unsupported interval visible and retain enough context to choose a more informative next observation. Preserve missingness flags and categorical vocabularies. Report the number and distribution of eligible, accepted, rejected, and recovered units under the same accounting rule. This prevents a clean-looking map or sequence from implying coverage it does not have. The substantive conclusion should concern reliability under the tested missingness conditions, with a clear account of how the fallback changes the scientific or operational target. An untested imputation strategy should not silently turn absent evidence into confident fact.

8.6 Deployment prevalence

The prevalence test for deployment prevalence takes an eligible flow stream and aggregated alert queue as its exposure population. The intended outcome is a manageable incident review workload. The difficulty is rare attacks and high traffic volume, which can make the evaluation mixture unlike the mixture encountered in use. State how the source sample was selected and which conditional rates are being transported to a new scenario. The independent reference is probability sampled reviewed observations and incident aggregation. A balanced development set can be appropriate for fitting, but its class proportions should not be mistaken for a measurement of the future population.

The proposed analysis is to evaluate declared prevalence and exposure scenarios before a live claim. Measure false alerts per time and confirmed incident fraction, retaining the eligible exposure denominator and the decision aggregation rule. A probability sample can estimate operating performance directly within its supported population. Scenario reweighting can expose consequences under alternative assumed proportions, but the assumed stability of conditional errors must be stated. It is often useful to report both the empirical source result and the scenario calculation. They answer different questions. Do not select only

periods with known target events when the claimed workload concerns a continuous stream that also contains long target-free intervals.

Interpret the result with the limit that balanced benchmark accuracy does not determine operational precision. Preserve traffic exposure and every alert aggregation rule, including the source sampling and any weights. A useful conclusion connects false detections and misses to a defined amount of observation, review, or action. It should show where uncertainty in prevalence materially changes the decision. Strong performance on an enriched collection can justify further study while remaining insufficient for a rare-event deployment claim. Making the population and exposure explicit helps readers distinguish those levels of evidence without dismissing the value of a carefully controlled benchmark.

8.7 Novel attack labels

The controlled perturbation in novel attack labels is applied to test records with labels absent from model fitting. The scientific target remains binary detection of a specified novelty subgroup; the factor varied is unknown attack categories and changed normal traffic. Begin by explaining why the intervention represents a meaningful change in the observation process. A numerical alteration can diagnose dependence without reproducing a real physical condition. Its value is greatest when the preserved information and the changed information are both explicit. The reference is the source label vocabulary and independent test labels, and any constructed component should retain a known design rather than being described as a new empirical observation.

The proposed experiment is to compare known label and absent label subgroups under one fixed threshold. Assess subgroup recall and normal false positives using the complete processing chain that the claim requires. If an intervention changes raw observations, recompute the actual derived inputs rather than changing a convenient proxy and assuming equivalence. Include a control that leaves the relevant physical quantity unchanged where that is possible. Compare errors across the declared perturbation range, avoiding a single unusually favorable setting. The test can reveal a dependence, but identifying its mechanism may require additional observations that vary the suspected factors independently.

The interpretive boundary is that binary novelty detection is not attack family identification. Preserve training vocabularies and subgroup membership, and keep the transformed observations distinguishable from the originals. A successful diagnostic explains which part of the model or measurement becomes unstable and at what support scale. It should not infer broad robustness from an algebraic property of one feature or an attractive transformed display. The next step can be a physical replication under the corresponding condition, with independent reference measurements. This progression connects a reproducible analytic test to the real observation without claiming that the former has already substituted for the latter.

8.8 Candidate promotion

The adaptation study for candidate promotion evaluates an incumbent and adapted candidate on later data. Its desired outcome is an independently supported replacement decision, and the main risk to interpretation is drift alarms without confirmed performance gains. A change detector and a model-quality comparison answer different questions. The former identifies a departure in a selected representation; the latter asks whether a candidate improves the intended output. Define the training cutoff, permitted labels, update budget, and promotion rule before the later evaluation. The independent reference is labels independent of candidate selection. Both an incumbent and a challenger should be assessed on the same eligible future observations.

The proposed comparison is to compare candidates on the same later labeled eligible stream. Evaluate paired misses false alerts and workload with paired records wherever possible. Labels acquired after an

update cannot be treated as if they had been available for fitting it. Pseudo-labels should retain their generated provenance and should not silently become independent reference truth. An annotation budget should include the actual selection and review process, because targeted easy examples can make adaptation look cheaper than it is. A frozen reference is valuable even when it is ultimately inferior, because it exposes whether change was needed and what the update actually contributed.

The interpretation must recognize that a distribution change does not prove a challenger is better. Preserve candidate versions cutoffs and promotion criteria, including rejected candidates and the conditions that triggered review. An adaptation result should report stability, coverage, and relevant errors rather than only the best post-update score. If the update is harmful for a rare subgroup, a pooled improvement can hide that consequence. A defensible promotion decision rests on independent later evidence and a declared operating objective. It should not be justified merely by a distribution alarm, a newer architecture, or a visually persuasive explanation of an individual prediction.

8.9 Incident decisions

The decision-level test for incident decisions uses an alert and its associated review context. It asks whether a defined response decision under stated costs remains supported after the model output enters the intended analysis or action. The factor of concern is correlated alerts and unequal asset consequences. Define the downstream quantity before examining model results, including any aggregation, threshold, matching, or cost assumptions. The reference is independent incident review and asset context. This creates a link between an intermediate score and the consequence that motivates the work. Without that link, a favorable technical metric can be assigned practical meaning that the experiment never measured.

The proposed comparison is to compare threshold and aggregation policies under declared cost structures. Report utility components and review burden alongside the intermediate errors, so that their relationship can be inspected rather than presumed. Different error types can have unequal effects on the endpoint. A small number of consequential errors may dominate a summary that looks stable on average. Conversely, some image or sample discrepancies may have little effect on the intended quantity. A sensitivity analysis should vary defensible decision assumptions without choosing the one that makes the model look best. Retain the underlying counts and source-supported measurements needed to assess alternatives.

The key limit is that a classifier flag alone is not an operational incident. Preserve underlying contingency counts and cost assumptions, including unresolved or declined decisions. A useful result identifies the conditions under which the model changes an endpoint, the uncertainty of that change, and the additional evidence required for a stronger interpretation. This is a substantive validation question, not an administrative final step. It allows a complex learning system to be judged on the scientific or operational contribution it actually makes while keeping broader claims available for separate, appropriately designed studies.

8.10 External traffic collection

The external study for external traffic collection concerns an independent network observation environment. Its claim is the declared transfer envelope for a detector, under changes in different protocols logging and normal behavior. Replication should begin from a locked feature, label, and decision specification, while documenting any necessary local mapping. The independent reference is local labels and verified field semantics. A new environment may use the same field names or category words while measuring something different. Check those meanings before comparing numerical scores. Unsupported mappings should remain visible so that a transfer result is not created by silently redefining the task.

The proposed experiment is to apply a locked feature contract and detector to a separately collected stream. Measure coverage calibration and condition specific errors, including the availability of valid inputs and the effort needed to obtain the reference. Report local conditions alongside the aggregate result. A model may be useful within a constrained envelope even when it fails elsewhere. The aim is to identify that envelope precisely, not to turn every local success into a universal claim. A reproducibility package should retain the configuration needed to reconstruct preprocessing, fitting, issue, and scoring, while preserving the permitted provenance of any source material that cannot be redistributed freely.

The limiting explanation is that historical benchmark performance is not a current network estimate. Preserve collection metadata feature mappings and complete predictions. The final account should distinguish what was observed, what was fitted, what was constructed for diagnosis, and what remains a proposed next test. Numerical reproducibility is necessary for the executed calculations, but it does not by itself establish external scientific validity. An informative replication can therefore confirm a useful result, reveal a boundary, or identify a measurement mismatch. Each outcome contributes knowledge when the source records and target definitions make the reason assessable.

9 Conclusions

The executed ridge-margin baseline reaches 0.76260 accuracy on the designated historical test file, with 0.63415 attack recall and 0.06766 false-positive rate. Recall falls to 0.44776 for attack rows carrying source labels absent from training. These results provide a transparent example of why pooled benchmark performance and detection of novelty should be examined separately. They are not current-network performance estimates.

Adaptation should proceed from a detected change to a diagnosed data condition and then to an independently evaluated candidate. The model's observation history, label timing, and alert policy must remain reconstructable. Combining those records with exposure-aware evaluation makes intrusion detection research more useful than optimizing a closed benchmark score alone.

References

[1] Zha, Chao, Kang, Zifeng, Liu, Tian, Shen, Dakun, Zhang, Ruyun, "Navigating the Latent Manifold: Proactive Concept Drift Adaptation for Resilient NIDS", arXiv preprint 2609.08623, 2026, [Online]. Available: https://arxiv.org/abs/2609.08623.

[2] Michel, Julien, Khan, Abdul Qadir, Jaber, Majed, Parrend, Pierre, "Concept drift mitigation through community and spectral graph analysis for the detection of cyberattacks in network traffic", arXiv preprint 2609.09442, 2026, [Online]. Available: https://arxiv.org/abs/2609.09442.

[3] Fernández-Barrios, Roberto, Pastor-López, Iker, Pikatza-Huerga, Amaia, Bringas, Pablo García, "Candidate Comparability Before Promotion: Conditional Validation in Adaptive Network Intrusion Detection", arXiv preprint 2609.04388, 2026, [Online]. Available: https://arxiv.org/abs/2609.04388.

[4] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, pp. 1-6, 2009, doi: 10.1109/cisda.2009.5356528.

[5] N. Moustafa, J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)", 2015 Military Communications and Information Systems Conference (MilCIS), pp. 1-6, 2015, doi: 10.1109/MilCIS.2015.7348942.

[6] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, A. Bouchachia, "A survey on concept drift adaptation", ACM Computing Surveys, vol. 46, no. 4, pp. 1-37, 2014, doi: 10.1145/2523813.

[7] S. Axelsson, "The base-rate fallacy and the difficulty of intrusion detection", ACM Transactions on Information and System Security, vol. 3, no. 3, pp. 186-205, 2000, doi: 10.1145/357830.357849.

[8] R. Sommer, V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection", 2010 IEEE Symposium on Security and Privacy, pp. 305-316, 2010, doi: 10.1109/sp.2010.25.

[9] L. Breiman, "Random Forests", Machine Learning, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[10] C. Cortes, V. Vapnik, "Support-vector networks", Machine Learning, vol. 20, no. 3, pp. 273-297, 1995, doi: 10.1007/BF00994018.

[11] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016, doi: 10.1109/CVPR.2016.90.

[12] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory", Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[13] A. Vaswani et al., "Attention Is All You Need", Advances in Neural Information Processing Systems, vol. 30, pp. 5998-6008, 2017, [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[14] G. E. Hinton, R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks", Science, vol. 313, no. 5786, pp. 504-507, 2006, doi: 10.1126/science.1127647.

[15] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, "Focal Loss for Dense Object Detection", 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999-3007, 2017, doi: 10.1109/ICCV.2017.324.

[16] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling", arXiv preprint arXiv:1803.01271, 2018, [Online]. Available: https://arxiv.org/abs/1803.01271.

[17] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. W. Vaughan, "A theory of learning from different domains", Machine Learning, vol. 79, no. 1-2, pp. 151-175, 2010, doi: 10.1007/s10994-009-5152-4.

[18] T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016, doi: 10.1145/2939672.2939785.

[19] D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure", Ecography, vol. 40, no. 8, pp. 913-929, 2017, doi: 10.1111/ecog.02881.

[20] C. Nadeau, Y. Bengio, "Inference for the Generalization Error", Machine Learning, vol. 52, no. 3, pp. 239-281, 2003, doi: 10.1023/a:1024068626366.

[21] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, "On Calibration of Modern Neural Networks", Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 1321-1330, 2017, [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html.

[22] T. Gneiting, A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation", Journal of the American Statistical Association, vol. 102, no. 477, pp. 359-378, 2007, doi: 10.1198/016214506000001437.

[23] B. Lakshminarayanan, A. Pritzel, C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles", Advances in Neural Information Processing Systems, vol. 30, pp. 6402-6413, 2017, [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

[24] S. M. Lundberg, S. I. Lee, "A Unified Approach to Interpreting Model Predictions", Advances in Neural Information Processing Systems, vol. 30, pp. 4765-4774, 2017, [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[25] R. Geirhos et al., "Shortcut learning in deep neural networks", Nature Machine Intelligence, vol. 2, no. 11, pp. 665-673, 2020, doi: 10.1038/s42256-020-00257-z.

[26] J. Bergstra, Y. Bengio, "Random Search for Hyper-Parameter Optimization", Journal of Machine Learning Research, vol. 13, pp. 281-305, 2012, [Online]. Available: https://jmlr.org/papers/v13/bergstra12a.html.

[27] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, "A Kernel Two-Sample Test", Journal of Machine Learning Research, vol. 13, pp. 723-773, 2012, [Online]. Available: https://jmlr.org/papers/v13/gretton12a.html.

[28] A. E. Hoerl, R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics, vol. 12, no. 1, pp. 55-67, 1970, doi: 10.1080/00401706.1970.10488634.