Albahit Journal of Applied Sciences
"open-access, peer-reviewed biannual journal"
eISSN: 2708-8936, pISSN: 2708-244X
Volume 5, Issue 4, 2026
Page No: 92-108
Website: *https://albahitjas.com.ly/index.php/albahit/en/index*



Calibration Allocation and Silent Decisions in Unseen Attack Intrusion Detection: A controlled empirical audit of reference support and threshold calibration

Ammar A. Ahmed [1*] , Enas Faris Yahya [2]
[1,2] Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq

[*]*Corresponding author: amrah@nahrainuniv.edu.iq*

Received: 28-07-2026	**Accepted:** 04-10-2026	**Published:** 18-10-2026
	Copyright: © 2026 by the authors. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract:
An intra-net intrusion detector may create a benchmark of benign samples to identify an alert threshold. We explored the tradeoff between these roles and how it impacts on attack families that are not included in model fitting. We took the fitted scores and evaluation records as fixed, and tested three different budget sizes and three different allocation shares over four different score constructions and three different input and calibration conditions. There were 4,860 related comparisons in grouped folds, and 12,374 withheld families and withheld seeds. There were no complete-input cases where an alert was issued by 405 empirical rank fusion cases. For 135 samples, the calibration sample was too small, and a finite threshold could not be found for the rule used. In the other 270, the threshold was the best possible. This ceiling effect was verified in 810 rank-fusion cases that have finite thresholds under input conditions. Some withheld families for both raw supervised scores and z-score fusion were alerted by scores from the same evaluation records. This diagnostic analysis used data from the original fitting domain. The results demonstrate the importance of considering the false-positive rates along with attack recall and the ability to exceed the alert threshold.

Keywords: intrusion detection; unseen attack families; missing features; false positive rate; score calibration; robustness audit.

تخصيص المعايرة والقرارات الصامتة في كشف التسلل للهجمات غير المرئية: تدقيق تجريبي مضبوط لدعم المراجع ومعايرة عتبات التنبيه

عمار عبد الرحمن أحمد[1*]، إيناس فارس يحيى [2]
[2،1] قسم علوم الحاسوب، كلية العلوم، جامعة النهرين، بغداد، العراق.

الملخص
قد يقوم نظام كشف التسلل داخل الشبكة بإنشاء مجموعة مرجعية من العينات السليمة لتحديد عتبة التنبيه. وقد استكشفنا المفاضلة بين هذين الدورين، وكيفية تأثيرها في عائلات الهجمات التي لم تُدرج ضمن عملية ملاءمة النموذج.

اعتمدنا درجات النموذج المُدرَّب وسجلات التقييم بوصفها معطيات ثابتة، واختبرنا ثلاثة أحجام مختلفة للميزانية وثلاث حصص مختلفة للتخصيص، عبر أربعة أساليب مختلفة لبناء الدرجات وثلاثة ظروف مختلفة للمدخلات والمعايرة.
أُجريت 4,860 مقارنة مترابطة ضمن طيات مقسمة إلى مجموعات، وشملت الدراسة 12,374 عائلة هجمات محجوبة وبذورًا محجوبة.
لم تُسجَّل أي حالات صدور تنبيه في حالات المدخلات المكتملة ضمن 405 حالات من حالات دمج الرتب التجريبي. وفي 135 حالة، كانت عينة المعايرة صغيرة للغاية، مما حال دون إمكانية تحديد عتبة محدودة وفقًا للقاعدة المستخدمة. أما في الحالات الـ270 المتبقية، فقد بلغت العتبة أفضل قيمة ممكنة.
وقد جرى التحقق من تأثير السقف هذا في 810 حالات من حالات دمج الرتب التي تمتلك عتبات محدودة في ظل ظروف المدخلات.
كما أُصدرت تنبيهات بشأن بعض عائلات الهجمات المحجوبة، سواء عند استخدام الدرجات الإشرافية الخام أو دمج درجات Z، وذلك استنادًا إلى الدرجات المستمدة من سجلات التقييم نفسها.
وقد استخدم هذا التحليل التشخيصي بيانات من نطاق التدريب الأصلي.
وتُظهر النتائج أهمية مراعاة معدلات الإيجابيات الكاذبة بالتوازي مع معدل استدعاء الهجمات، والقدرة على تجاوز عتبة التنبيه.

الكلمات المفتاحية: كشف التسلل؛ عائلات الهجمات غير المرئية؛ السمات المفقودة؛ معدل الإيجابيات الكاذبة؛ معايرة الدرجات؛ تدقيق المتانة.

Introduction

An alerting rule can go almost “invisible” and therefore generate very few false alarms and not see malicious traffic. When evaluating the detector on attacks not used for model fitting, this elementary distinction is important. The correct evaluation should report the benign false positive rate and the excludedfamily recall rate. None of these measures (accuracy or false alarm count) is sufficient on its own to provide detection coverage of any use.

A second concern is related to the use of auxiliary networks for observations. Development may generate a service-history counter or application metadata that isn't provided on a subsequent input to a model. When the imputation values are replaced by fitted ones, the scores distribution changes. Next, there can be a calibration update which changes the operating point and also in a fuse detector, the ordering of records. Only telling someone a score before and after a recalibration is performed masks what causes the score.

The main issue is whether a fixed total benign-record budget can result in various operational failures when divided up into branch-reference construction and final threshold calibration. We fix detector parameters and evaluation records and investigate that allocation varying, and whether a finite rank-ceiling saturation or the selected tolerance rule fails to yield a finite threshold. Confounding factors are provided by the auxiliary-history removal, separate reference and threshold refresh, and raw-score controls and prior model-training controls. There is no claim of the elementary ceiling identity or the existence of a conservative calibration.

Related evidence and originality boundary

Yue, Lu and Xie [1] have already implemented supervised and benign-deviation evidence, aligned the heterogeneous scores by benign references, and set a final low-FPR threshold. Our standard detector has the same general ingredients and doesn't appear as MCDE or as a new alternative with algorithmic priority. It is used to reveal the behaviour of an existing style of detector when subjected to a controlled observation intervention.

Di et al. [2] investigate corruption aware training of network intrusion detection with random masking, block masking and noise. Replicated-clean training controls are also described in their publication abstract. Our simple masked-copy control utilizes trees, where each row added is a random choice from a set of pre-defined blocks, and no proportional noise. This is an

inspired default, not a reproduction of MaskAug. Limiting full-text access means they cannot claim it doesn't contain a specific analysis.

Outside IDS, availability-conditioned calibration is also set up. In the study of missing scans and conditional conformal prediction by Ajit [3] valid but uninformative outputs are also considered. Azizi [4] derives availability-aware calibration for multimodal regression. Those targets are not excluded-family binary alerting, but it's not enough novelty in changing the application. Ahmed et al. [5] report on auditing misleading label-free monitoring in temporal IDS shift and Arepalli et al. [6] report on family-dependent novelty and conformal triage. In a non-peer-reviewed preprint, Biswas [7] explores the protocol and drift effects. The general caution that calibration or a comforting monitor may go hand in hand with inadequate discrimination is nothing new.

The empirical contribution adopted is a fixed total budget allocation audit for operationally different failure causes, fixed raw score controls and documented execution trail. Discrete calibration resolution and zero discoveries have already been studied by Hennhoefer and Preisach [8]. Abdulwahab et al. [9] have dealt with the provisioning, contamination and subsequent chronological recalibration in the calibration-window archive, which does not justify the use of it as evidence that calibration size is a new topic. The research by Szabo and colleagues [10] is on ECDF score fusion and false-positive certificates. The narrower allocation intervention and its measured intervention failure map was not found as an equivalent experiment in the primary material that was inspected. This bounded distinction does not imply great algorithmic novelties or a first ever mathematical claim, it's just moderate empirical novelty. Limitation of literature access and coverage still exist.

Material and methods

Data and evaluation populations

According to the documentation at UNSW, there are 175,341 rows in a training partition and 82,332 rows in a testing partition [11]. Those are the dimensions of our downloaded mirrors, and the number of predictor fields we expected: 42, and we have two targets, and id. Matching is not enough to define the identity of the bytes as they are not official schema or count identity. The train.csv was taken from Hugging Face's training version of the CSV from Mouwiya/UNSW-NB15, while the test version was taken from Mireu-Lab/UNSW-NB15 under the weird name train.csv. The dimensions to be audited are identified by local file names. The result manifest contains hashes of SHA-256.

We remove the id and both the targets from the predictors, and remove the sttl, dttl and the TTL-derived ct_state_ttl. This gives a view with 39 predictors. The following are purposively masked: ct_srv_src, ct_srv_dst, ct_dst_ltm, ct_dst_src_ltm and ct_src_ltm. The service-history mask takes off the first two, the endpoint-history mask takes off the next three, and the combined mask takes off all five. This is a human man-made feature availability intervention. It is not a network outage or intentional traffic alteration by an attacker that has been observed.

Table 1. Data populations and calibration support

Population	Rows	Purpose
Training mirror	175341	Role allocation
Testing mirror	82332	Confirmation
Source confirmation	15732	Previously unscored role
Test after base embargo	61481	Sensitivity population
Benign reference	2185	Branch remapping
Benign threshold set	2092	Final operating point

The 34 field base view is used to hash into roles, omitting all five masked counters and the three TTL related fields. 100% base profile matches remain in their source profile. About 60% of the deterministic first hash is allocated to fitting, 15% allocated to development selection, 15% allocated to calibration and 10% allocated to confirmation. Two other deterministic hashes divide calibration into two benign sub-sets, reference and threshold, which are preserved from the previous experiments. The role of source is never used for tuning the final rule. The categories were previously labeled in descriptive counts, but was described as "not totally unseen in every sense" as opposed to being described as "unseen.

The 20,851 rows that are excluded from the testing sensitivity are the 34-field base view that occurs anywhere in the source mirror. The intent of this, conservative, source-wide embargo is to cast a wider net than excluding records that are actually in one fold. It's not the earlier 8,541-row full-feature overlap count, since the grouping view is different. All test results and embargoed test results are still available. The population will change if overlap is removed and the removal of overlap should not be considered a pure leakage correction at fixed test composition.

Both supply CSV, but none of these have timestamps. Grouping by exact-profiles does not create host-disjoint, session-disjoint or chronological evaluation. Different flows can have a representation that is the same, and there are feature groups in the source with different binary and/or category labels. These will not be automatically designated as annotation errors, but will be included in the as-distributed analysis. Withheld-family membership doesn't imply new real world exploit – it implies category removal from fitting.

Modeling and decision interventions

For supervised branch, it is either Random Forest [12] or ExtraTrees [13] with 60 number of trees, maximum depth of 16, and minimum leaf size of 3. There are 30,000 source-fit records which are sampled by each family-specific fit, with one of the three fits excluded at a time. The anomaly branch is a 150-tree Isolation Forest [14] fitted on and at most 30,000 source-fit benign records using max_samples=2048. Control sampling, masking & fitting at 7 and 42 seeds 2026. Only records are used to learn numeric medians, categorical modes and one-hot dictionaries. Categories that are not found in the data are ignored by the fitted one-hot encoder. Clean training, doubled clean training and mask-augmented training are compared. The latter two consist of a copy of each row and the original rows. The augmented copy gets one of the three predefined masks, in a uniform random manner. Labels are duplicated in a uniform way. This reduces the number of rows to be scanned and the number of duplicate rows returned, but not exactly the time of execution, the tree structure, or the information content returned. Within each training regime both branches, including imputers, are refitted.

Let $s_j(x)$ denote the attack-oriented score of branch j, and A_j its benign reference scores. The empirical map is:

$$r_j(x) = \frac{\left|\{a \in A_j : a < s_j(x)\}\right|}{|A_j| + 1}$$

We evaluate $\max_j r_j(x)$, the supervised map alone, and the anomaly map alone. The strict comparison and its tie handling are fixed in code. Individual branch remapping is monotone but can introduce ties; two separate branch maps followed by max fusion may alter cross-record ranking. Branch-only controls help distinguish fusion behavior from threshold effects.

A separate benign set is used to provide a strict score $>$ threshold alert rule. If there is no eligible rank, the threshold is ∞; otherwise, it is the smallest eligible order statistic for which:

$$\text{BinomialTail}(n, 0.99, k) \leq 0.05$$

Given a fixed score and independent identically distributed benign calibration and deployment records, there is a nominal 1% tail bound at a 95% confidence level. Those assumptions are not necessarily made with the flow data and shifted/masked conditions. Thus, we report empirical FPR, rather than a certified population or drift-proof bound. AUROC and standardized partial AUROC up to FPR $= 0.01$ are diagnostics that measure label-aware ranking, not independently calibrated operating recall.

Table 2. Fixed-model decision states for one missing-input pattern.

State	Branch references	Threshold
Complete baseline	Complete input	Complete input
Missing frozen	Complete input	Complete input
Reference only	Pattern matched	Complete input
Threshold only	Complete input	Recalibrated with old maps
Fully recalibrated	Pattern matched	Recalibrated with new maps

In a training program, the family, seed and backbone, detector parameters and evaluation records are not varied through Table 2.Three controlled contrasts are separated by the ordered complete-to-missing frozen → reference-only → fully-recalibrated path. They are simple telescoping algebra, and nothing new or unique about causal explanations. An alternate update order is available in the threshold-only corner. For a metric M, the interaction is:

$$\Delta_{\text{int}}(M) = M(\text{full recalibration}) - M(\text{reference only}) - M(\text{threshold only}) + M(\text{frozen})$$

If there is a nonzero interaction ($\Delta_{\text{int}}(M) \neq 0$), then an order-independent allocation is not suitable.

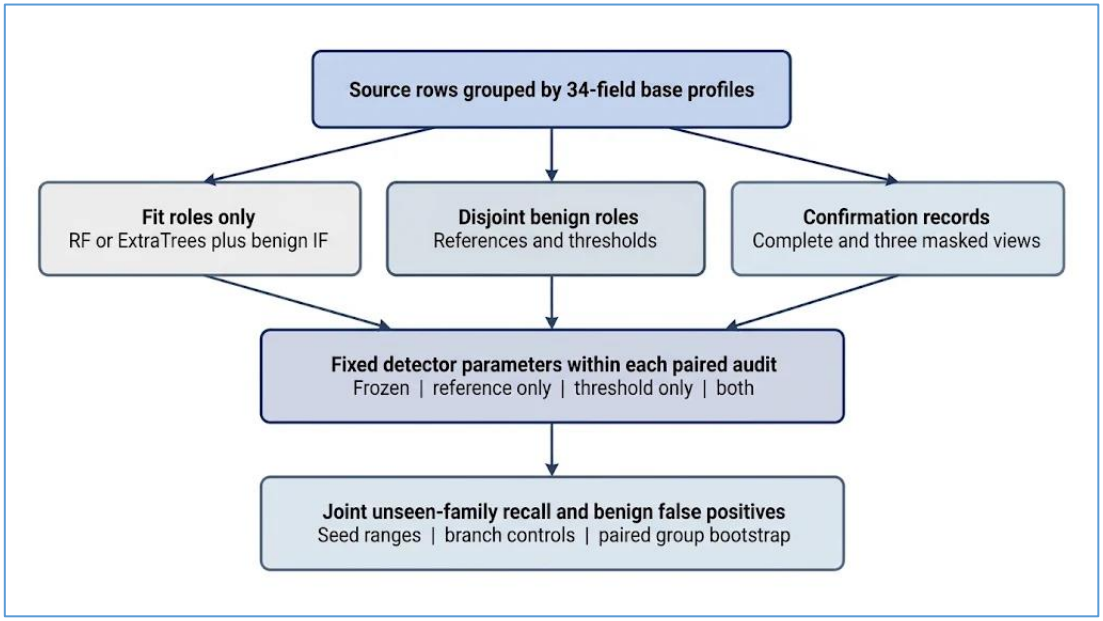

Figure 1. Audit workflow with disjoint role allocation and paired decision interventions.

Confirmation protocol and uncertainty

The final protocol JSON and script hash were saved prior to scoring for confirmation. Following that evaluation is no final-label threshold adjustment, model selection or parameter search. All the development and TON results are marked as development evidence. There are 7,776 metric records in the locked grid, comprised of 3 seeds, 3 training regimes, 3 excluded families, 2 supervised backbones, 3 evaluation populations, 4 input masks (including

complete), 3 score rules and 4 calibration corners. These are not 7,776 independent experiments since corners are repetitive and results are recorded and shared.
The input equality check was not complete in the first attempt to score confirmation and was subsequently rejected. Small scores differences occurred as a result of parallel tree-probability summation, which had an impact on strict ties. In protocol revision 2, the scoring process is modified to score one worker, repeated calls to calibration are added to a cache, and bit identical reference replays are asserted prior to evaluation. The failed outputs will be saved and not included in all tables and figures reported. None of those failed outcomes led to the selection of a hyperparameter, family, training regime or threshold. This is an externally unpreregistered untouched computational correction.
We refer to means and ranges for seeds but do not consider them separate sampling events. The seed-2026 clean RF fusion resamples the exact base-profile groups of each class 1,000 times. Each resampled group maintains their row multiplicity and the same group of draws applies to complete and missing predictions. The intervals so obtained represent the conditional sensitivity of grouped resampling. They fail to validate independent network events, nor do they give confidence at the same time over the experiment grid. Hard to find categories and large repeats can result in limited effective observations.

Results and discussion

Confirmation results

Table 3. Clean RF fusion confirmation means over three fitting seeds.

Population	Excluded family	Complete R / F	Missing frozen R / F	Missing recal R / F
Source	Reconnaissance	45.26 / 0.39	13.11 / 0.20	13.47 / 0.37
Source	Shellcode	50.90 / 0.55	9.04 / 0.31	14.47 / 0.30
Source	Generic	37.32 / 0.55	11.58 / 0.37	17.30 / 0.36
Embargoed test	Reconnaissance	28.19 / 0.65	2.42 / 0.22	2.88 / 0.70
Embargoed test	Shellcode	47.17 / 0.93	4.54 / 0.51	7.14 / 0.55
Embargoed test	Generic	65.55 / 0.92	19.37 / 0.58	21.63 / 0.67

R is excluded-family recall and F is benign FPR, both percentages. Means compare the same records within each population and seed. The embargoed population is a sensitivity analysis with different composition from the full test mirror. Figure 2 adds fitting-seed ranges; these whiskers are not independent-split confidence intervals.

Table 3a. Source confirmation mask sensitivity with matched calibration and clean RF fusion

Family	Complete R / F	Service R / F	Endpoint R / F	All R / F
Reconnaissance	45.26 / 0.39	23.49 / 0.39	37.63 / 0.43	13.47 / 0.37
Shellcode	50.90 / 0.55	20.93 / 0.39	24.29 / 0.31	14.47 / 0.30
Generic	37.32 / 0.55	24.33 / 0.46	11.94 / 0.41	17.30 / 0.36

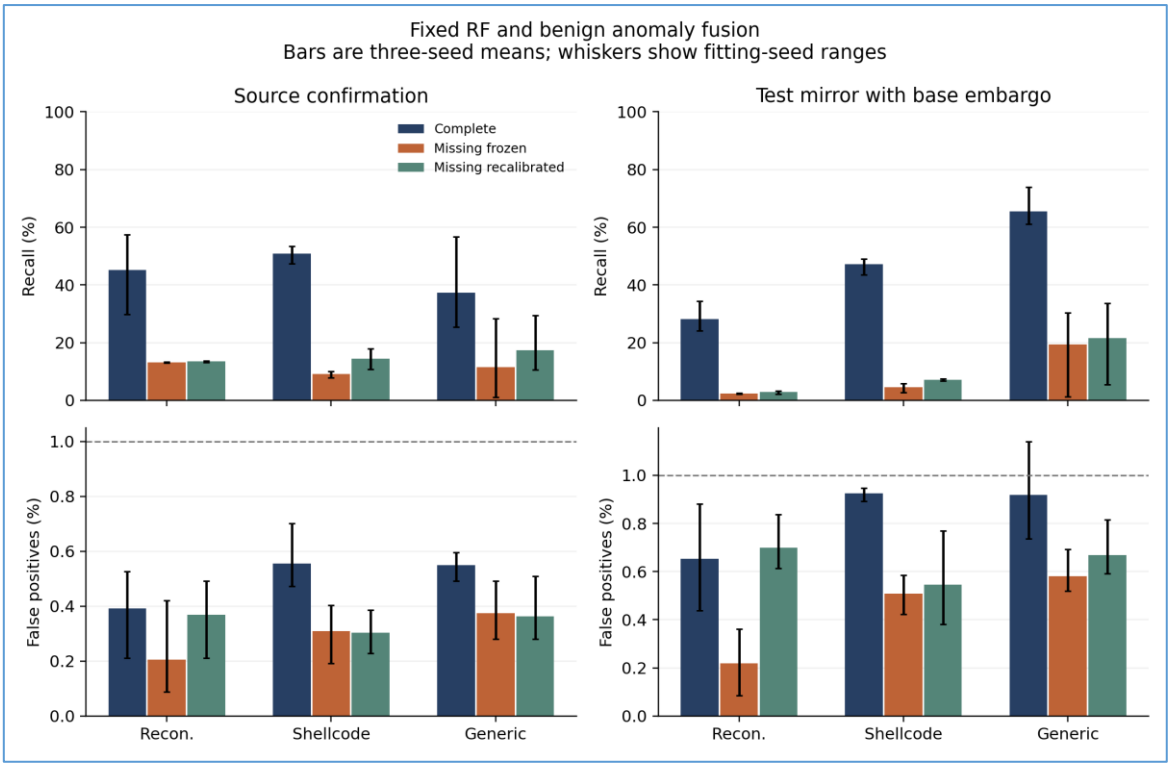

Figure 2. Joint recall and false positive behavior with complete versus absent history fields.

When comparing the mean value of clean-fusion family by backbone by population, it is found that in 18 of the 27 cases, the missing-input recall is lower than its complete counterpart. In 15 of those cases the FPR is at or below 1% (missing-input mean). These are just a summary of related conditions and not hypothesis tests nor estimates of the frequency of the effect in deployment. The machine readable supplement contains all-test results, branch-only results and partial-AUROC diagnostics.

Training and backbone controls

Table 4. Source confirmation under combined history loss and full recalibration.

Backbone	Family	Clean R / F	Doubled clean R / F	Mask augmented R / F
RF	Reconnaissance	13.47 / 0.37	14.21 / 0.45	19.38 / 0.45
RF	Shellcode	14.47 / 0.30	6.98 / 0.29	9.30 / 0.46
RF	Generic	17.30 / 0.36	24.77 / 0.43	7.34 / 0.49
ExtraTrees	Reconnaissance	23.46 / 0.47	14.76 / 0.94	24.27 / 0.40
ExtraTrees	Shellcode	21.96 / 0.67	20.93 / 0.51	13.44 / 0.74
ExtraTrees	Generic	3.95 / 0.78	5.09 / 0.31	22.65 / 0.24

The doubled-clean control tests whether simply providing repeated training records explains an augmentation comparison. The complete-input results of each refitted regime are also available: a masked-condition improvement must be weighed against any complete-input deterioration. Differences between Random Forest and ExtraTrees are backbone sensitivity, not independent datasets, and the Isolation Forest component is shared in construction.

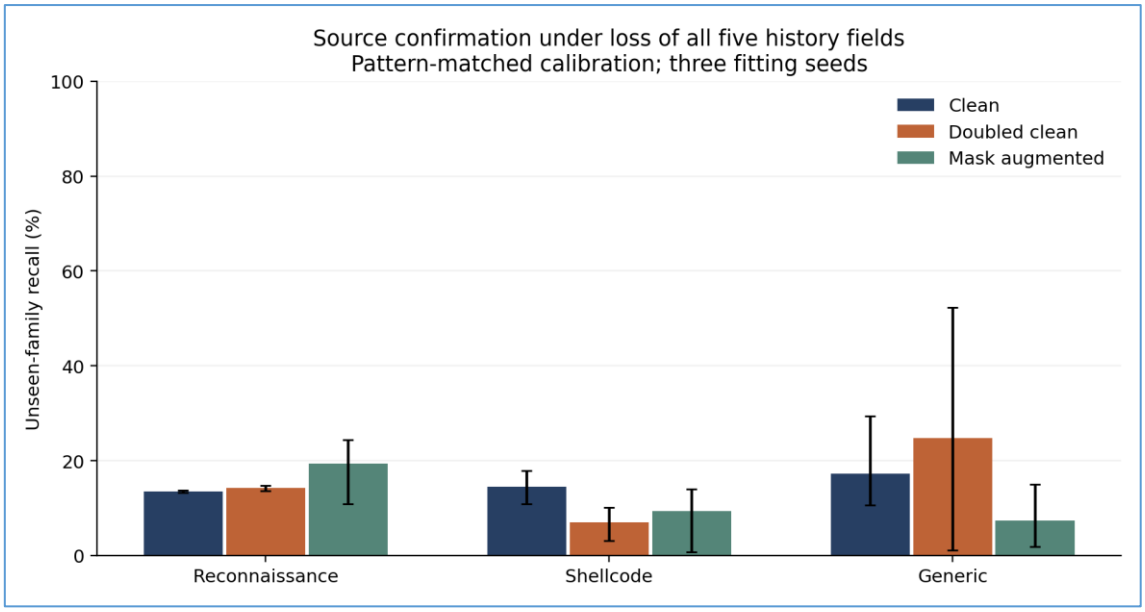

Figure 3. Training controls with identical source confirmation records and pattern-matched calibration.

Table 5. Branch-only source confirmation means for the RF training grid

Family	Score	Complete R / F	Missing recal R / F
Reconnaissance	Supervised	58.19 / 0.42	41.12 / 0.24
Reconnaissance	Anomaly	0.11 / 0.54	0.00 / 0.43
Shellcode	Supervised	61.76 / 0.36	21.45 / 0.30
Shellcode	Anomaly	0.00 / 0.54	0.00 / 0.43
Generic	Supervised	4.66 / 0.57	19.57 / 0.33
Generic	Anomaly	48.09 / 0.54	0.72 / 0.43

The anomaly-only fit does not utilize the excluded family label, which means that the underlying model is shared by family comparisons within a seed and regime, but that each family defines a different attack evaluation subset. These rows should not be perceived as three separate anomaly detectors that are fitted independently. Despite a high AUROC, a poor recall at the chosen low FPR level can still be allowed, which is why operating metrics are still important.

Paired uncertainty and update order

Table 6. Paired group bootstrap for source confirmation recall changes

Family	Missing rule	Groups	Change pp	95 percent interval pp
Reconnaissance	Frozen	603	-44.05	-62.19 to -24.16
Reconnaissance	Recalibrated	603	-43.94	-61.41 to -24.07
Shellcode	Frozen	115	-45.74	-55.08 to -36.43
Shellcode	Recalibrated	115	-42.64	-52.42 to -33.58
Generic	Frozen	190	-28.50	-58.82 to 3.84
Generic	Recalibrated	190	-27.41	-58.89 to 4.03

Intervals in Table 6 are only valid for seed 2026 using clean RF fusion, assuming the fitted models, calibration and observed source group population. Changes in percentages are

expressed by comparison of missing with full input. These do not contain uncertainty due to a new calibration sample, uncertainty due to family of models, uncertainty due to arbitrary missingness, or uncertainty due to traffic drift. Test-population intervals and FPR intervals are included in final_group_bootstrap.json.

Table 7. Reference by threshold interaction for the seed 2026 source audit.

Family	Recall interaction pp	FPR interaction pp
Reconnaissance	-0.11	-0.33
Shellcode	-12.40	-0.33
Generic	-1.05	-0.02

Even with a small or zero interaction, the interaction is reported. If not zero, the assignment of the total calibration change between reference refresh and threshold refresh will vary based on the path selected. This does not imply causation between events in the real network. A reference refresh updates the maps with the same thresholds, while a threshold-only update is calibrated separately with the old maps; the four corners of this design are the explicit executable rules.

Native Schema Replication on TON_IoT

The TON_IoT network mirror includes 211,043 records (50,000 normal records and nine attack labels) from codymlewis/TON_IoT_network at revision ed5bc29125a5d057be27b8b251ec3170adaa9cf4. Official corpus completeness is yet to be confirmed.

- **Features included:** The native model uses 11 flow/service (svc) predictors and 21 DNS, TLS, HTTP, and anomaly-metadata fields.
- **Features excluded:** It does not include direct addresses, ports, raw queries, URIs, user agents, or certificate identities.
- **Deterministic missingness:** Missingness in affected fields is deterministically caused by invalid values of http_trans_depth.
- **Dataset context:** The TON_IoT collection corresponds to the same underlying data as a larger collection described in the source dataset documentation and the original network-dataset publication.

[15].

TON features native-schema models that are fitted using a base-profile hash split of around 60% fitting, 15% benign reference, 15% benign threshold and 10% evaluation. To exclude ddos, injection or password, use the outer RF fit. With the stress, all of the 21 application metadata fields are removed. These are a set of distinct variables to UNSW history counters. This is not direct model transfer or an identical input intervention, but rather, is an input

intervention that is corpus-specific. Tone outcomes are not new confirmatory evidence, but rather are checked throughout development.

Table 8. TON native schema development replication means over three fitting seeds.

Family	Condition	Recall percent	FPR percent
ddos	Complete	73.13	0.66
ddos	Missing frozen	72.75	0.95
ddos	Missing recalibrated	70.03	0.39
injection	Complete	75.50	0.66
injection	Missing frozen	71.50	1.07
injection	Missing recalibrated	34.75	0.46
password	Complete	41.98	0.62
password	Missing frozen	40.69	0.91
password	Missing recalibrated	23.59	0.41

Matched recalibration significantly decreases the FPR of these TON families, with relatively small changes in recall, whereas the frozen condition in which the metadata is lost causes significant decreases in injection and password recall. This means that the UNSW interpretation of auxiliary loss does not apply to all scenarios, and recalibration does not apply to all scenarios as a panacea for better attack coverage. It is not a question of whether the calibration is good or bad in general, but rather a question of whether the joint recall and false-alarm outcome, under a given input pattern, is good or bad.

Grouped partition sensitivity

An extra fixed-protocol sensitivity analysis is conducted on the original source fitting role only, without all the source-confirmation and testing record. The three outer folds get the same 34-field grouping of base-profile, but a different fixed hash. They do not have common members in their evaluation groups. For each outer training pool, 70% of the groups are allocated to the fitting pool, 10% of groups to the benign score reference pool and 20% of groups to the benign threshold pool, each with a fixed assignment via hash. The previously selected model settings, seed 2026, training controls and masks do not change. Training populations are not the same for different folds, this is not three separately collected sets of data or newly acquired confirmation.

Table 9. Grouped partition sensitivity means within the original fitting domain.

Family	Training	Complete R / F	Missing frozen R / F	Missing recal R / F
Reconnaissance	Clean	0.00 / 0.00	0.00 / 0.00	33.95 / 0.60
Reconnaissance	Doubled Clean	0.00 / 0.00	0.00 / 0.00	30.17 / 0.57
Reconnaissance	Mask augmented	5.40 / 0.30	3.69 / 0.04	20.98 / 0.65
Shellcode	Clean	0.00 / 0.00	0.00 / 0.00	18.31 / 0.51
Shellcode	Doubled Clean	0.00 / 0.00	0.00 / 0.00	20.35 / 0.64
Shellcode	Mask augmented	13.64 / 0.32	3.41 / 0.04	17.67 / 0.65
Generic	Clean	0.00 / 0.00	0.00 / 0.00	17.01 / 0.57
Generic	Doubled Clean	0.00 / 0.00	0.00 / 0.00	11.92 / 0.63
Generic	Mask augmented	7.17 / 0.30	0.46 / 0.03	10.82 / 0.70

Not a simple missing-input collapse claim can be made on the partition extension. Complete-input fusion has already been achieved without noise in all three folds, and the doubled-clean version also does not emit any noise. The fitted threshold is 0.9995693367786391, which is the same as the empirical-map ceiling 2321 / 2322. Of all the benign records that fall within the threshold, only 94 exceed the threshold. Note that the score for alerts needs to be greater than the threshold, so that there is no possible input that can lead to the activation of this particular bounded rule. It makes all the corresponding TP-FP count pairs to be reproduced exactly in the replay. This is not a new theorem, or a statement that all calibrated IDSs including MCDE break this way, it is just a diagnosed saturation mechanism for the implemented rule.
Unweighted means shown in Table 9 are based on three evaluation supports. All 432 condition results are retained with respect to the fold. These folds were only added after the initial confirmation stage, and are therefore a sensitivity extension and not a part of the original frozen grid. Before aggregation the following are checked for complete-input invariance, role-group separation, paired sample supports, script hashes and count identities. The rows that make up the final data set are not altered; the original 7,776 metric rows are maintained. Fold ranges are descriptive, and the design cannot claim a three independent datasets significance.
Score mapping controls
A second diagnostic corrects the original clean RF score and benign IF score in the same three original-fitting-domains, only changing the way in which they are constructed. It is a comparison between empirical-rank max fusion, raw supervised probability, raw anomaly score and maximum z-score fusion. The last control is performed with the benign reference mean and the benign reference standard deviation of the branch, with a modified standard deviation in the case it is 1e-12 or lower. The benign threshold protocol is the same as usual, but disjoint. The z-score control brings out the empirical-rank ceiling, but does not imply that it is immune to other kinds of ceiling and ties/branch imbalance. A normalization process before score fusion is done [16]. They are typical score controls and not a new detector nor an exact duplicate of that biometric study.

Table 10. Fixed score construction controls within the three grouped folds

Family	Score construction	Complete R / F	Missing frozen R / F	Missing recal R / F
Reconnaissance	Rank max	0.00 / 0.00	0.00 / 0.00	33.95 / 0.60
Reconnaissance	Raw supervised	65.83 / 0.71	46.01 / 0.24	55.38 / 0.68
Reconnaissance	Raw anomaly	0.03 / 0.28	0.03 / 0.02	0.20 / 0.60
Reconnaissance	Z score max	0.03 / 0.28	0.03 / 0.02	5.47 / 0.60
Shellcode	Rank max	0.00 / 0.00	0.00 / 0.00	18.31 / 0.51
Shellcode	Raw supervised	72.15 / 0.88	25.24 / 0.38	33.25 / 0.74
Shellcode	Raw anomaly	0.00 / 0.28	0.00 / 0.02	0.00 / 0.60
Shellcode	Z score max	0.00 / 0.28	0.00 / 0.02	0.00 / 0.60
Generic	Rank max	0.00 / 0.00	0.00 / 0.00	17.01 / 0.57
Generic	Raw supervised	2.97 / 0.80	16.08 / 0.70	28.54 / 0.79
Generic	Raw anomaly	15.54 / 0.28	0.01 / 0.02	16.18 / 0.60
Generic	Z score max	15.54 / 0.28	0.01 / 0.02	16.18 / 0.60

The raw supervised control is significantly superior for the complete-input recall for Reconnaissance and Shellcode, and is still low for Generic. Maximum z-score fusion does not give a uniform remedy and can recall the weak anomaly branch. The opposite is true, however, if the history fields are removed – the recall will actually get worse. There are 144 rank-control TP and FP pairs that exactly replay Stage 15. As expected, raw-branch reference-only updates do not affect their scores. All the additional 576 metric rows and raw score caches are

preserved. These comparisons are based on previously eaten diagnostic folds and have no ability to choose and self-certify a fresh final deployment rule.
Established the total allocation for calibration costs.Set a total budget for calibration.
The outcomes of the allocation protocol are determined prior to the protocol. Total benign-record budgets are 512, 1024 and 2048, and reference shares are 0.25, 0.50 and 0.75. Five subsampling seeds are 11, 29, 47, 71 and 101. They already have roles as reference and threshold records in their existing disjoint sources. Fold and seed shuffles maintains matched record choices across methods within a fold and seed. All comparisons are made using the same RF/IF parameters and evaluation records. Equating record counts is a modelled verification budget or not a measurement of collection time or physical telemetry cost. There are 4860 related rows of metrics where the input is complete, missing history is frozen calibration, and missing history is matched recalibration. No new model fitting or scoring of the closed original confirmation populations results.

Table 11. Complete input rank fusion failure causes under fixed total record budgets.

Total	Reference share	Reference records	Threshold records	Cause in all 45 cases
512	0.25	128	384	Finite ceiling saturation
512	0.50	256	256	Insufficient threshold support
512	0.75	384	128	Insufficient threshold support
1024	0.25	256	768	Finite ceiling saturation
1024	0.50	512	512	Finite ceiling saturation
1024	0.75	768	256	Insufficient threshold support
2048	0.25	512	1536	Finite ceiling saturation
2048	0.50	1024	1024	Finite ceiling saturation
2048	0.75	1536	512	Finite ceiling saturation

The 405 complete-input rank-fusion cases in Table 11 are: 135 cases of insufficient-support fallback for the selected tolerance construction, and 270 cases with finite ceilingsaturated thresholds. No case has complete input rank emits alert. This is no silence in these limited source pools if the tested overall budget is increased 4 times. These are observed rules (not general rules) for these rules, pools and partitions, not for all IDS models, not for all allocations. All 1620 cases of insufficient support, 715 cases of rank ceiling, 265 other finite rules with no alerts seen and 2260 other finite rules with alerts seen are still present in the full grid. These totals are not the number of trials or the probability of deployment failure but rather the number of cells related.
The map implemented and maximum fusion value are ceiling $c = n/(n+1)$ for n references per each branch. The final finite threshold will equal c if and only if the number of calibration scores not exceeding c is at least m-k+1; if not, then final finite threshold will be larger than c. If final finite threshold is larger than c, then no valid input will alert. All cases of finite-support rank, there are 810, are checked for this elementary identity and it is not a new theorem. Finites that do not produce any alerts on the observed records are placed in a different category since there is no way to infer structural impossibility from an observed 0 count.
Discussion and limitations
The allocation audit demonstrates the reasons why the reachability and support for calibration must be separated prior to interpreting a low false positive. The paired design separates score degradation from an update of the decision rule but does not assert any unique causal contributions of the components. One of the added folds reveals this as well, in that the "all inputs present" baseline should be evaluated for saturation before interpreting missing-input contrasts, since a "silent" baseline is incapable of demonstrating additional missing-input recall

collapse. The estimates shown here are based on raw scores, which test that some observed silence is directly related to the way the scores are constructed, whereas the findings based on simple standardization and branch fusion do not solve all families. Also a negative result can be useful: an augmentation that works for one family may be detrimental to another, and bad false positives don't help. A deployment assessment should maintain the input-availability condition, consider the benign support, attack coverage simultaneously, and not use the same evaluation label to determine the value of the threshold.

There are multiple generalisations limits. There are only three UNSW families and three TON families to be examined. The three UNSW masks are all subsets of the same five counters. Two sample tree backbone models are not a representative survey of modern neural or sequential IDS models. Near-duplicate profiles and shared hosts may not be matched and there are no timestamps. The test mirror is a second one party of the same corpus and not an independent institution or modern live network. Changes in embargo sensitivity change the test population and could selectively exclude common traffic patterns.

Development employed a number of previous diagnostics, while this is a final protocol, which only freezes the final confirmation stage and is not externally preregistered. The newly coupled evaluation phase introduces a split in evaluation domain coverage aspects of partition sensitivity but the training sets are overlapping and do not create independent acquisition-domain replication. The simple augmented-copy control doesn't reproduce MaskAug, nor does it establish that missingness aware training is generally bad. The full-data model objects were not stored during the early stages; the deterministic refits were compared to the previous TP and FP counts. Final prediction caches for main seed-2026 clean-RF audit, complete metrics / executable refitting code for larger grid.

The closest-literature review comes to establishing an empirical distinction, rather than a definitive statement about priority. Full protocol exclusion is not possible when there is restricted access to one close article. For a strong publication, it may be necessary to use exact baseline reproductions and more independent acquisition domains, and split-level replication, as well as richer missingness mechanisms. Such extensions without solutions do not warrant changing of the observed outcomes or presenting an ordinary calibration as a new invention.

Conclusion

The study makes a contribution in the form of a controlled empirical audit of the division of a fixed benign-record budget between the number of score references and the number of final threshold calibrations. In the cases of measured rank-fusion, the resulting silence is visible by two different mechanisms—lack of finite-threshold support and finite score-ceiling saturation. This silence is not likely to be confused as a reliable unknown-attack detection, or as a universal information-loss effect, due to fixed-score, missing-input and branch controls. The originality is experimental intervention, measured allocation failure map, cause specific analysis are established models, calibration tools and mathematical ceiling identity. The results are limited to their source pools, the ways they are grouped, and the timing of their explorations; they do not contain a new detector, a new statistical guarantee, absolute priority or deployment utility.

Reproducibility and declarations

The package contains pinned Python requirements, dataset audit manifests and hashes, fixed role definitions, and the final protocol (frozen), fitting scripts, full JSON metrics, paired prediction caches, grouped bootstrap results, figures and previous development results. When the experiments are run, they won't overwrite their output files. To make a clean copy: install requirements, check hashes for data, execute the fitting scripts in the prescribed order, and then

execute the scripts summarize_final.py and build_research_report.py. Report and final plots are created using results stored, without re-running closed fitting stages.
There are no reported competing interests and no external funding was reported for this project. The experiments will examine public benchmark mirrors and not attack operational networks. No new human-participant data are gathered. Literature discovery, code preparation, execution of experiments and report drafting were done by AI assistance. The researcher is responsible for checking the evidence, approving the authorship and implementing the eventual venue's disclosure policy. No claims of unsupported self- or institution-affirmed authorship and ethics are made.

References

[1] J. Yue, Y. Lu and Y. Xie. Detecting Unseen IoT Attacks with Calibrated Dual Evidence Under Low False-Positive Budget. Entropy, 28(9), 1026, 2026. DOI 10.3390/e28091026.

[2] Q. Di, K. Du, G. Chen and H. Deng. MaskAug: Architecture-agnostic corruption-aware training for robust network intrusion detection. Journal of Computer Security, published online 14 September 2026. DOI 10.1177/0926227X261487917.

[3] A. Ajit. Guarantees That Survive a Missing Scan: Modality-Conditional Conformal Prediction for Multimodal Medical Diagnosis. MLMI, MICCAI satellite paper, 2026. https://papers.miccai.org/miccai-2026-sat/paper/MLMI_024.pdf

[4] I. Azizi. Conformal Calibration for Multi-Modal Regression with Missing Modalities. Proceedings of Machine Learning Research, 329, 343-377, 2026. https://proceedings.mlr.press/v329/azizi26a.html

[5] W. Ahmed, M. A. Rahaman, P. Roy and P. Alam. The dashboard says perfect: a systematic audit of label-free model monitoring and test-time adaptation under temporal distribution shift in network intrusion detection. Frontiers in Artificial Intelligence, 9, 1928858, 2026. DOI 10.3389/frai.2026.1928858.

[6] S. V. Arepalli, U. S. G. Irrinki, A. Morampudi and R. K. Surapani. Conformal prediction for network intrusion triage and analysis of unseen attack families. Discover Informatics, 1, 33, 2026. DOI 10.1007/s44564-026-00033-y.

[7] M. A. Biswas. The Evaluation Protocol Is the Hidden Variable: A Drift-Aware Re-Examination of Flow-Based Intrusion Detection Benchmarks. Preprints.org, version 1, posted 11 June 2026. Not peer reviewed. https://www.preprints.org/manuscript/202606.0903

[8] O. Hennhoefer and C. Preisach. Between Resolution Collapse and Variance Inflation: Weighted Conformal Anomaly Detection in Low-Data Regimes. arXiv:2603.23205, submitted 24 March 2026. Preprint. https://arxiv.org/abs/2603.23205

[9] S. Abdulwahab, M. Z. Abdullah and A. Sallomi. Provisioning label-free alert budgets: measurement code and results. Zenodo, version 1.1, 2026. DOI 10.5281/zenodo.22081860. The current successor record, version 41, concerns chronological recalibration: DOI 10.5281/zenodo.23468304. These are software and output archives, not peer-review certification of the present study.

[10] L. Szabo, Z. Weltsch and A. Adamne-Major. When Does Score Fusion Help? Conformally Certified Out-of-Distribution Detection for Camera and LiDAR Sensors. Sensors, 26(15), 4706, 2026. DOI 10.3390/s26154706.

[11] N. Moustafa and J. Slay. UNSW-NB15: a comprehensive data set for network intrusion detection systems. Military Communications and Information Systems Conference, 2015. DOI 10.1109/MilCIS.2015.7348942. Dataset documentation: https://research.unsw.edu.au/projects/unsw-nb15-dataset

[12] L. Breiman. Random Forests. Machine Learning, 45, 5-32, 2001. DOI 10.1023/A:1010933404324.

[13] P. Geurts, D. Ernst and L. Wehenkel. Extremely randomized trees. Machine Learning, 63, 3-42, 2006. DOI 10.1007/s10994-006-6226-1.
[14] F. T. Liu, K. M. Ting and Z.-H. Zhou. Isolation Forest. IEEE International Conference on Data Mining, 413-422, 2008. DOI 10.1109/ICDM.2008.17.
[15] N. Moustafa. A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. Sustainable Cities and Society, 72, 102994, 2021. DOI 10.1016/j.scs.2021.102994. Dataset documentation: https://research.unsw.edu.au/projects/toniot-datasets
[16] A. Jain, K. Nandakumar and A. Ross. Score normalization in multimodal biometric systems. Pattern Recognition, 38(12), 2270-2285, 2005. DOI 10.1016/j.patcog.2005.01.012.

The detailed procedure for the experiments is given in Appendix A.

This appendix outlines the way the workflow was performed in Sections 3–10. The main confirmation grid and the subsequent (partition diagnostics) and fixed score allocation audit outputs are different and address different questions. This extended description does not bring any further model fitting.

A1 Planning the work and giving records to others

We loaded the 175,341 row training-sized mirror and the 82,332 row testing-sized mirror, verified the field structure and got the file hashes. Removed attack_cat, id and label from candidate predictors. We reduced sttl, dttl and ct_state_ttl as well, so now there are 39 input fields. The labels were kept for fitting and evaluation separately. The numeric medians, categorical modes and one-hot vocabularies were fit to the records for the specific roles and used for all subsequent ones. The fitted encoder ignored categorical values that were unknown.

The grouping representation was created by dropping the five history fields from the 39 field view. This 34-field profile used hashing to ensure that exact matches were within a source role. The estimated percentages were fitting (60 percent), selection (15 percent), calibration (15 percent), and confirmation (10 percent). Calibration was split up into separate benign reference and benign threshold functions. It's only a deterministic partition, but the percentages indicate hash allocation and not an exact row-count target. The number of roles assigned and the source-row identifiers are included in the saved audit outputs.

Only profiles from the source mirror that were not in the middle of the mirror were excluded in the testing sensitivity (20,851 test rows left 61,481). It is a conservative test in which the test population is changed. These checks are not about representation independence in time or between institutions, as there is no way to get a host-disjoint partition or timestamps.

A2 Identifying unspecified families and missing inputs

We excluded each family from the supervised fitting pool and kept other attacks and benign traffic. We repeated the procedure for Reconnaissance, Shellcode and Generic. Only fitting records were used for isolation forest. Evaluation was then performed on the withheld family, versus with benign records. This category withholding protocol is not an exploit or a performance test against any unknown attack.

Table A1 Executed feature availability interventions

Condition	Fields replaced by missing values
Service history	ct_srv_src, ct_srv_dst
Endpoint history	ct_dst_ltm, ct_dst_src_ltm, ct_src_ltm
All history	All five fields above
Complete	No deliberate history-field removal

The imputed missing entries were done by the fitted preprocessing Pipeline. During the evaluation-time interventions, the detector parameters remained fixed. This is an isolated event where no predictors were recorded, not an adversarial event where traffic was measured to be absent.

A3 Fitting and making alert decisions

Random Forest and ExtraTrees both had 60 trees, a depth of most 16, and a minimum leaf size of 3; with 30,000 trees sampled per excluded-family fit the number of eligible fitting rows. The benign-only Isolation Forest took 150 trees, max_samples 2,048 and at most 30,000 fitting rows. The main repeated fits were controlled by Seeds 2026 and 42. The corrected final scoring implementation used one worker to maintain comparisons in a deterministic way when scores were tied, but was not used by Fitting.

Samples records were used only once for clean training. When doubled cleaning was used, the original records were duplicated. Mask-augmented training was used to add one copy, where one of the three pre-defined history

masks was sampled uniformly per copied row. No labels were changed and pre-processing and estimators were re-fit for each regime. The duplicated-clean control is different than the previous method of increasing the number of training rows because it also provides a masked-copy comparison.
Suitable reference mapping and threshold determination.
Both branches were oriented to enhance the evidence of attack; the higher the score, the more evidence of attack. For branch j, a benign reference sample A_j defines r_j(x) = count{a in A_j : a < s_j(x)} / (|A_j| + 1). Important: The strict less than rule is enabled – tied scores will be mapped to the same value. The largest branch rank is used for maximum fusion. Each branch was evaluated independently as well. The raw supervised probability was later used as one score control, as well as raw anomaly scores and maximum z-scores fusion with reference means and population standard deviations.
The final threshold is provided by a second benign sample. If there are n threshold records, then the following tolerance rule is implemented which selects the smallest ascending rank k such that P(Binomial(n, 0.99) >= k) <= 0.05 and takes the kth score. If none of the ranks qualify, then the threshold is infinite. A record will only alert if its score is above the threshold. This is a nominal 1 percent tail target at 95 percent confidence, assuming the usual iid benign assumptions of fixed scores, but, because of empirical dependence of flows and changes in flow distributions, this cannot be used as a proven deployment guarantee.
Evaluated four corners of the calibration when one or more inputs were not present: keep both original components; update branch references only; update original maps to update the final threshold only; or update both components. All corners have the same fitted detectors and evaluation rows. All corners are exactly compared for complete-input repeated corners. This will allow us to see the contribution of changing scores and changing decision calibration to the measured difference.
A5 Main confirmation grid measuring.
This produces 7,776 main metric rows which are the result of 3 fitting seeds, 3 training regimes, 3 excluded families, 2 supervised backbones, 3 evaluation populations, 4 input conditions, 3 score rules and 4 calibration corners. Evaluation populations are source confirmation, the full testing mirror and its source-profile-embargoed subset. These are not independent trials since they share the same records and have repeated conditions.
True positives were counted in the excluded-family records and false positives were counted in the benign records. Recall (TP): number of evaluated withheld-family attacks/total number of withheld-family attacks.FPR (FP): number of evaluated benign records/total number of benign records. AUROC and standardized partial AUROC up to FPR 0.01 is included in the ranking summaries. The means and ranges are descriptive of seed. For some specific seed-2026 comparison for clean-RF, base-profile groups are sampled from each class and multiplicities are respected in 1,000 paired bootstrap resamples.1,000 paired bootstrap resamples sample base-profile groups from each class to preserve multiplicities of groups for selected seed-2026 clean-RF comparisons. These intervals are based on the observed group population and fitted models.
A6 Diagnosing silence and allocating calibration records
The latter diagnostic folds are only applied to the original fitting domain. Base-profile grouping created three outer evaluation folds that are not connected to each other. A fixed hash value, which is about 70 percent used for fitting, 10 percent is used for benign references and 20 percent is used for benign thresholds, was used in each of the remaining training pools. Training pools are overlapping in folds. The 432 partition-sensitivity results are then an evaluation of the sensitivity to splitting in one acquisition source. They are not three datasets that are independent of each other.
Then we kept the RF and IF scores constant, and compared different constructions of these scores. The 144 exact TP and FP replays of the rank controls are in the 576 score-control rows. This set that some complete-input rank-fusion rules were silent, while others were useful with some of the families of raw-supervised rankings. The score caches were used as the inputs for the allocation experiment and no population or closed source-confirmation was rescored.

Table A2 Fixed-score allocation experiment factors

Factor	Executed settings
Total benign budget	512, 1,024, 2,048 records
Reference share	0.25, 0.50, 0.75
Subsampling seeds	11, 29, 47, 71, 101
Grouped folds and families	3 folds and 3 excluded families
Score constructions	Rank max, raw supervised, raw anomaly, z-score max
Conditions	Complete; missing frozen; missing recalibrated

We extracted qB records as references and the rest for thresholds from the existing disjoint roles for each budget B and reference share q. The allocation comparison was paired with the allocation of shuffled prefixes, which were used in all methods. 4860 rows of metric data were generated by crossing the Table (A2). Equal record counts represent a verification budget, collection time and physical telemetry costs were not measured. These

allocations outcomes were made in the context of a plan that was already established prior to these outcomes, but the source data had already been used to inform earlier diagnostics.

A7 Assigning the reason for a silent rule

We first examined the possibility of finding a finite threshold for the selected tolerance construction. We did not report adequate threshold support if it could not. The maximum possible fused score for finite empirical-rank rules with n equally sized branch references is c = n/(n+1). If the threshold is done with strict alert comparison, it means that at least m-k+1 of the m calibration values will be equal to c, which will not alert. We checked this elementary ceiling condition in 810 finite-support cases. It is not a new theorem, but rather an existing identity of order-statistics.

There are no input ranks that fuse completely: 135 of them were not supported by enough thresholds, and 270 had a finite saturation ceiling. At all methods and conditions, there were 1,620 rows lacking sufficient support, 715 rows with rank-ceiling saturation, 265 rows with other finite rules observed with no alerts, and 2,260 finite rules observed that triggered an alert. A finite rule was taken to be separate from a mathematically unattainable threshold, if it occurred in the data, while that rule having zero alerts was maintained. The counts refer to related experiment cells, rather than probabilities of deployment failures.

A8 Keep evidence and understand the contribution

Each reported result is tied to executable code via stored plans, script and data hashes, raw score caches, per-condition metrics and integrity checks. The invalid parallel-scoring run will be stored separately and won't be included. There is a read-only verifier to check the main grid, partition and score-control replays, and allocation outputs. The contribution is defined as the allocated intervention, quantified separation of failure mechanisms. No winning allocation was chosen by evaluating labels, and the evaluation of these exploratory diagnostics are not sufficient to establish a new operating rule.

The example of 512-record allocation is a good instance. The assignment of 25 per cent to references gives 128 reference records and 384 threshold records and all 45 cases of complete input rank have finite ceiling-saturated thresholds. Assign 50 percent to give 256 records per role; assign 75 percent to leave 128 thresholds records. In all 45 of the latter cases, both of these allocations trigger the insufficient-support fallback of the chosen rule. Although the outcome of the alert is the same (visible zero), the alerting mechanism is different. To interpret an operating point in these measured pools, an increase in a total budget is not enough.

Table A3 Execution outputs and their evidential roles

File or directory	Retained evidence
final_validation.json	7,776 valid main metric rows and frozen protocol
stage15_partitions.json	432 grouped partition sensitivity rows
stage16_score_controls.json	576 fixed-model score-construction controls
stage16_scores/	Raw scores used by the allocation audit
stage18_plan.json	Budget allocation plan and executed script hash
stage18_allocation_audit.json	4,860 allocation metric rows
stage18_integrity_checks.json	Grid, budget, metric and ceiling checks
verify_checkpoint.py	Read-only checkpoint verification

To view the finished practical work, start with the 'per-condition' metric files, not re-run fitted models. Match condition of experiment with its family, fold or seed, input mask, then read TP, FP and their respective denominators. Make comparisons before aggregation. The reproducibility guide explains how to run the program in a clean copy, overwrite guards prevent completed units from being overwritten. There are no new performance observations, and this expanded appendix provides documentation of existing work.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of ALBAHIT and/or the editor(s). ALBAHIT and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content