



WHEN COST-SENSITIVE THRESHOLDS FAIL TO TRANSFER: A LEAKAGE-CONTROLLED EVALUATION OF NETWORK INTRUSION DETECTION ON UNSW-NB15

¹CHANDRA PRAKASH MISHRA²ATUL KUMAR MISHRA

¹M.Tech Research Scholar; ²Assistant Professor

Computer Science and Engineering, Millennium Institute of Technology and Science,
Bhopal, MP, India

chandraprakashmishragwl@gmail.com

Abstract: Cost-sensitive intrusion detection seeks to prevent missed attacks without overwhelming analysts with false alarms. However, selecting a threshold on held-out development data does not guarantee that its operational advantage will survive a change in traffic composition. This paper examines that transfer problem using the official UNSW-NB15 partitions (175,341 training and 82,332 testing records). A leakage-controlled workflow separates model fitting, probability calibration, and threshold selection within the training partition, and holds the official test partition untouched until evaluation. Four conventional baselines are compared with a Platt-calibrated random-forest policy using an illustrative missed-attack-to-false-alarm cost ratio of 5:1. The random forest achieves 89.42% test accuracy and an F1-score of 0.9105. Calibration with a validation-selected threshold of 0.13 reduces false negatives from 1,009 to 17, but increases false positives from 7,700 to 15,458. Thus, despite a substantial validation-set cost reduction, observed test cost rises from 0.1548 to 0.1888 per flow (21.95%); the paired-bootstrap 95% interval for the increase is [0.0298, 0.0381]. Three completed seed-wise partition checks show the same direction. The reported behavior coincides with an attack-prevalence change from 68.06% in training to 55.06% in testing, without establishing causation. Our results support evaluating threshold transfer, calibration stability, and decision costs alongside accuracy, rather than treating recall gains as evidence of lower operational risk.

Index Terms— Network intrusion detection; UNSW-NB15; cost-sensitive classification; probability calibration; threshold transfer; dataset shift; random forest.

I. INTRODUCTION

Network intrusion detection systems (NIDSs) help security operators identify traffic that warrants investigation. A classifier's utility depends not only on separating malicious and benign observations but also on the errors its decisions create for the operator. A false negative can leave an attack unnoticed; a false positive can trigger triage work, ticket generation, or interruption of legitimate service. These outcomes have different consequences, yet much of the benchmark literature foregrounds accuracy or F1-score without an explicit decision policy [1, 2].

A common adjustment is to lower an attack-probability threshold so that more flows receive alerts. The resulting increase in sensitivity is predictable; whether it reduces an assumed operational cost is an empirical question. If attack prevalence, feature statistics, or score calibration changes between threshold selection and subsequent testing, a decision rule that looks favorable on development data can become expensive. Security-learning evaluation is especially vulnerable to misleading comparisons when preprocessing leaks information, splits are not independent, or dataset construction differs from deployment conditions [3, 4].

This paper uses a fixed, public benchmark to study a narrower and falsifiable question: Does a calibration- and cost-based threshold selected entirely on development data improve realized cost on the independently provided UNSW-NB15 test partition? We intentionally distinguish the quality of probability ranking from the consequences of a particular classification threshold. The objective is not to introduce a new random-forest architecture or claim a general counterexample to cost-sensitive learning. Rather, the contribution is an experimentally grounded account of a threshold-transfer failure under an explicit, reproducible split discipline.

The work makes three specific contributions. First, it separates feature fitting, probability calibration, threshold selection and final evaluation while comparing simple models on identical official partitions. Second, it reports the confusion counts and realized asymmetric costs that reveal a recall–false-alarm trade-off obscured by a single headline metric. Third, it adds paired-bootstrap uncertainty, three completed seed-wise development-split checks, and a fixed-decision sensitivity analysis

to identify where the cost preference changes. All model scores and figures are derived from executed experiments; external published results are discussed as related literature, not treated as directly comparable test outcomes.

II. RELATED WORK AND RESEARCH POSITIONING

2.1 Benchmark-Based Intrusion Detection

Moustafa and Slay introduced UNSW-NB15 to combine benign and simulated malicious traffic with a broad family of connection-level attributes [5]. Its official portal supplies both larger source files and a predefined training/testing partition [6]. The partition is convenient for comparable experiments, but public benchmark performance cannot by itself establish performance on live enterprise traffic. Ring et al. discuss the representativeness and construction limitations of network intrusion datasets [4]; Arp et al. identify general security-ML measurement pitfalls, including problematic data partitions and misleading generalization claims [3].

Kasongo and Sun studied UNSW-NB15 with XGBoost-informed feature selection followed by classifiers including logistic regression, k-nearest neighbors, decision trees, support-vector machines and neural networks [7]. Their study motivates feature-space and classification comparisons, but it does not supply a controlled estimate of the cost change between the exact policies evaluated here. We do not reproduce their feature-selection procedure, and their published accuracies are not used as interchangeable baselines for our preprocessing or splits.

2.2 Calibration, Thresholds and Decision Cost

A score with good ranking can still be unreliable as a posterior probability. Probability calibration seeks agreement between predicted probabilities and observed frequencies; approaches include sigmoid/Platt scaling [8, 9]. We evaluate both the Brier score [10] and an exploratory ten-bin expected calibration error (ECE) on the locked test partition. Neither ranking metrics nor validation-set calibration error alone proves that a model's decisions are robust to dataset changes.

Recent work by Cha et al. optimizes cost-sensitive, attack-class-specific thresholds using XGBoost on intrusion benchmarks [11]. Their experiments span multiple seeds and datasets, whereas the present study evaluates a binary global threshold for a calibrated random forest on one official train/test partition. The methods, cost ratios and operating points differ; no cross-paper improvement claim follows from comparing their reported percentages with ours. The substantive distinction here is examination of the development-to-official-test failure and the associated cost break-even boundary.

III. RESEARCH METHODOLOGY

3.1 Official Partitions and Feature Boundary

We use the two user-provided CSV files matching the official UNSW-NB15 train/test record counts [6]: 175,341 training flows and 82,332 test flows. Each provided file has 45 columns, including a binary label, an attack-family descriptor `attack_cat`, and a row identifier `id`. The binary target is one for an attack and zero for benign traffic. The identifier, target, and attack-category field are excluded from predictor variables. The category field is retained only for post-test descriptive analysis; learning from it would risk direct label leakage. The workflow also excludes `difficulty_level` if present.

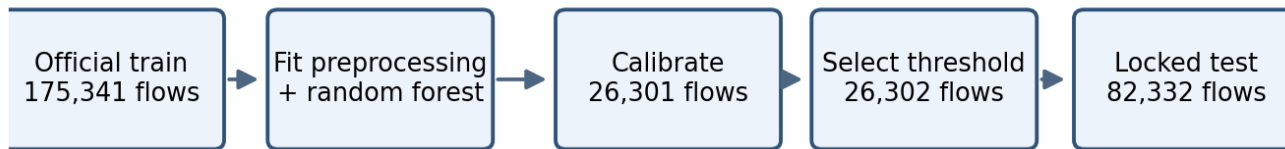
Table 1: Official dataset partitions and internal development split

Subset	Records	Attack share
Official training	175,341	68.06%
Model development	122,738	stratified
Calibration	26,301	stratified
Threshold selection	26,302	stratified
Official testing	82,332	55.06%

The first split allocates 70% of the official training records to fitting and 30% to an intermediate pool; the pool is halved for independent calibration and threshold selection (random seed 42). Stratification retains the official training label proportions. This is a record-wise split; flow dependence and possible shared origins mean it should not be mistaken for host-, episode-, or time-disjoint verification. The test partition is neither used to construct features nor to choose the threshold.

Numerical features receive training-only median imputation and robust scaling. Categorical attributes receive training-only most-frequent imputation and one-hot encoding with unknown categories ignored. A pipeline refits these transformations within each model's development data. Infinite numeric values are converted to missing values beforehand. No balancing or feature statistics are computed on the test partition.

All learned transformations and thresholds are fitted without test labels



Within official training split: 122,738 development / 26,301 calibration / 26,302 selection

Fig. 1: Data and decision boundary. All model fitting and policy selection remain inside the official training set; the locked official test split is used only for evaluation.

3.2 Model Specification and Calibration

The comparisons are a class-prior dummy classifier, a class-weighted logistic regression (500 maximum iterations), a class-balanced random forest (95 trees, minimum leaf size 3, square-root feature subsampling), and histogram gradient boosting (110 iterations, 18 maximum leaf nodes, learning rate 0.08, and L2 regularization = 1). The project uses scikit-learn [12] with reproducible random states; the paper's method is not a deep neural model. Random forests are a conventional nonlinear reference [13].

For the random forest, sigmoid calibration is fitted to an independent calibration subset. With a raw attack score $s(x)$, the sigmoid calibration mapping has the form

$$q(x) = 1 / \{1 + \exp[-(a \times s(x) + b)]\} \quad (1)$$

where a and b are estimated using calibration-only labels and $q(x)$ is the calibrated attack probability. Calibration adjusts score interpretation; it does not create a new ranking model.

3.3 Decision Rule, Metrics and Hypothesis

For true binary labels y and an attack decision y_{hat} , define the illustrative mean cost per flow as

$$\text{Mean cost} = [C_{\text{FN}} \times \text{FN} + C_{\text{FP}} \times \text{FP}] / N \quad (2)$$

The policy sets a false-negative cost of 5 and a false-positive cost of 1. This 5:1 ratio is an analysis assumption, not an independently measured monetary cost. We evaluate 189 candidate thresholds uniformly spanning 0.03 to 0.97, selecting the minimum validation cost; ties prefer the smaller threshold. The selected calibrated threshold for seed 42 is 0.13. Conventional baselines use 0.5. The risk comparison is undertaken on the locked official test partition after selection. Different values examined later are expressly post-hoc sensitivity analyses of fixed decisions, not test-optimized thresholds.

We report accuracy, precision, attack recall, binary F1-score, average precision (AP), area under the receiver-operating-characteristic curve (ROC AUC) [2, 14], Brier score, confusion counts, and realized cost. For a paired comparison of two policies on the same test records, we bootstrap the per-record cost difference with 1,500 class-stratified resamples at a fixed test-class mix; its percentile 95% interval describes uncertainty under a simplifying independent-flow assumption. These resamples do not estimate uncertainty from time shift or new attack environments.

IV. RESULTS AND DISCUSSION

4.1 Classification Performance on the Locked Test Split

Table 2 reports all five policies on 82,332 test records. The default random forest has the highest observed accuracy (89.42%) and F1-score (0.9105) among the evaluated operating points. Histogram boosting attains a slightly higher AP (0.9886 versus 0.9881) but lower F1 at the specified threshold. Calibration and threshold selection do not change the random forest's discrimination ranking (AP remains 0.9881). They do change the positive classification rate and thereby the precision, recall and realized error costs.

Table 2: Binary-classification performance on the official UNSW-NB15 test set (seed 42)

Classifier / policy	Acc.	Prec.	Recall	F1	Bal. acc.	AP	ROC AUC	Brier
Class-prior baseline	.5506	.5506	1.0000	.7102	.5000	.5506	.5000	.2643
Logistic regression	.7053	.6821	.8704	.7649	.6867	.8374	.8026	.1753
Random forest ($\tau=.50$)	.8942	.8520	.9777	.9105	.8848	.9881	.9841	.0675
Histogram gradient boosting	.8791	.8280	.9850	.8997	.8671	.9886	.9847	.0793
Calibrated RF ($\tau=.13$)	.8120	.7456	.9996	.8542	.7909	.9881	.9841	.0818

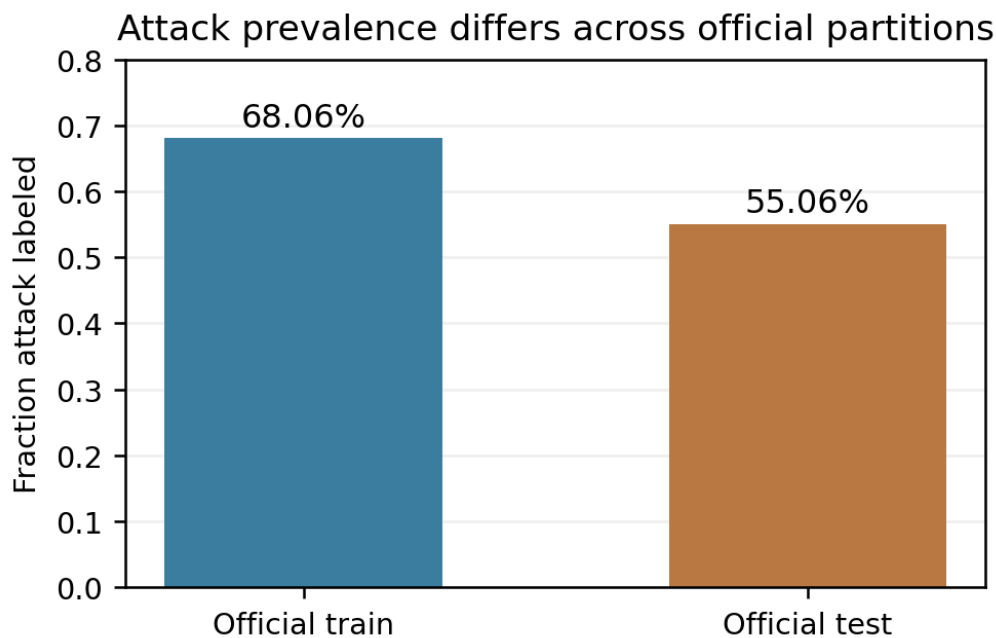


Fig. 2: Observed attack prevalence in the official provided partitions; the difference establishes a change in class proportions, not its causal mechanism.

4.2 Calibration and Threshold Transfer

On the internal threshold-selection subset, the default forest has a 5:1 cost of 0.1499 per flow; the calibrated, selected decision lowers it to 0.0638. On the official test set, the same comparison reverses: 0.1548 for the default forest versus 0.1888 for the selected policy (Fig. 3). This is a 0.0340 absolute and 21.95% relative increase in cost, despite improved attack recall. The stratified paired-bootstrap 95% interval for the cost difference is [0.0298, 0.0381] per flow. It does not include zero under the sample-level resampling assumptions.

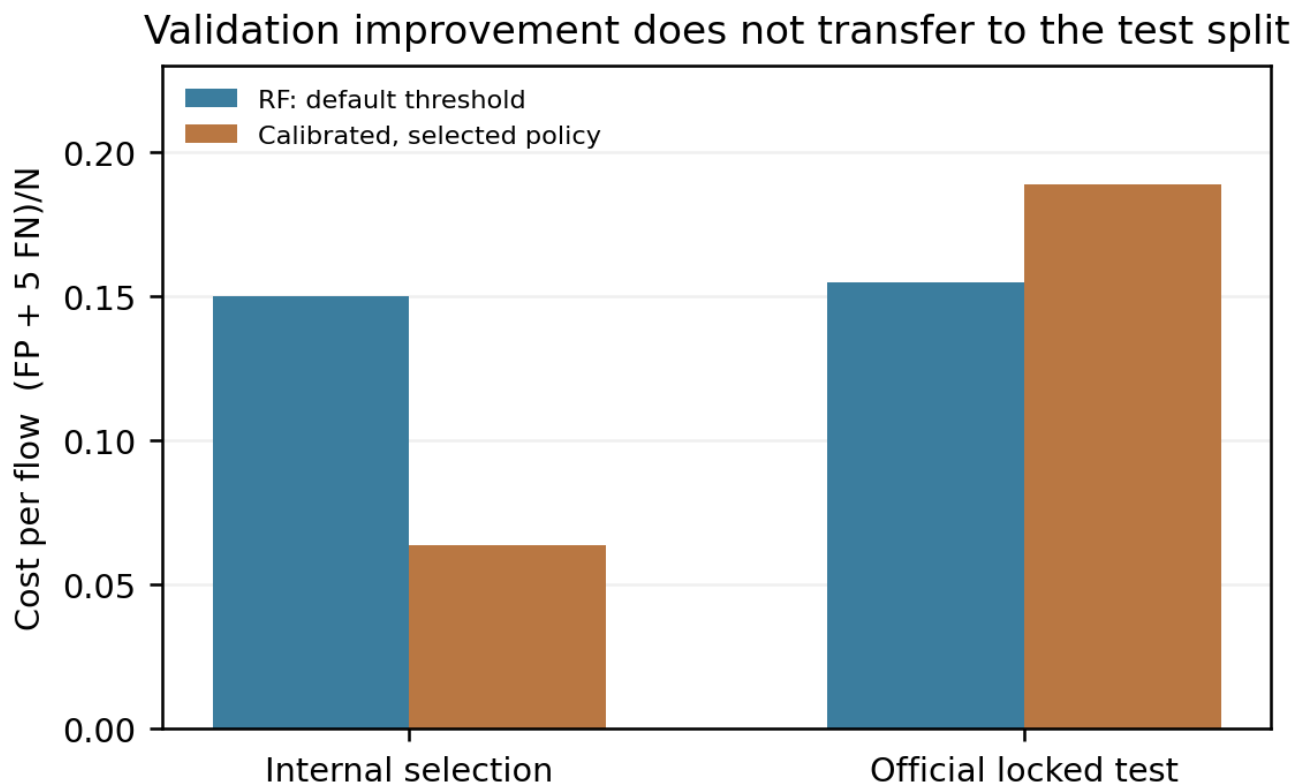


Fig. 3: Observed five-to-one weighted error cost on selection and locked-test subsets. Test values were not available during threshold selection.

The underlying confusion counts provide a direct accounting (Table 3). Lowering the decision threshold identifies 992 additional attacks previously missed, but generates 7,758 additional benign-flow alerts. The default forest has $(7,700 + 5 \times 1,009) / 82,332 = 0.1548$ cost per flow. The calibrated policy has $(15,458 + 5 \times 17) / 82,332 = 0.1888$. The attack-detection gain is real, but it is not equivalent to a lower cost at the stated ratio.

Table 3: Test-set confusion counts and cost comparison

Quantity	Default RF	Calibrated policy
True negatives	29,300	21,542
False positives	7,700	15,458
False negatives	1,009	17
True positives	44,323	45,315
5:1 weighted cost / flow	0.1548	0.1888

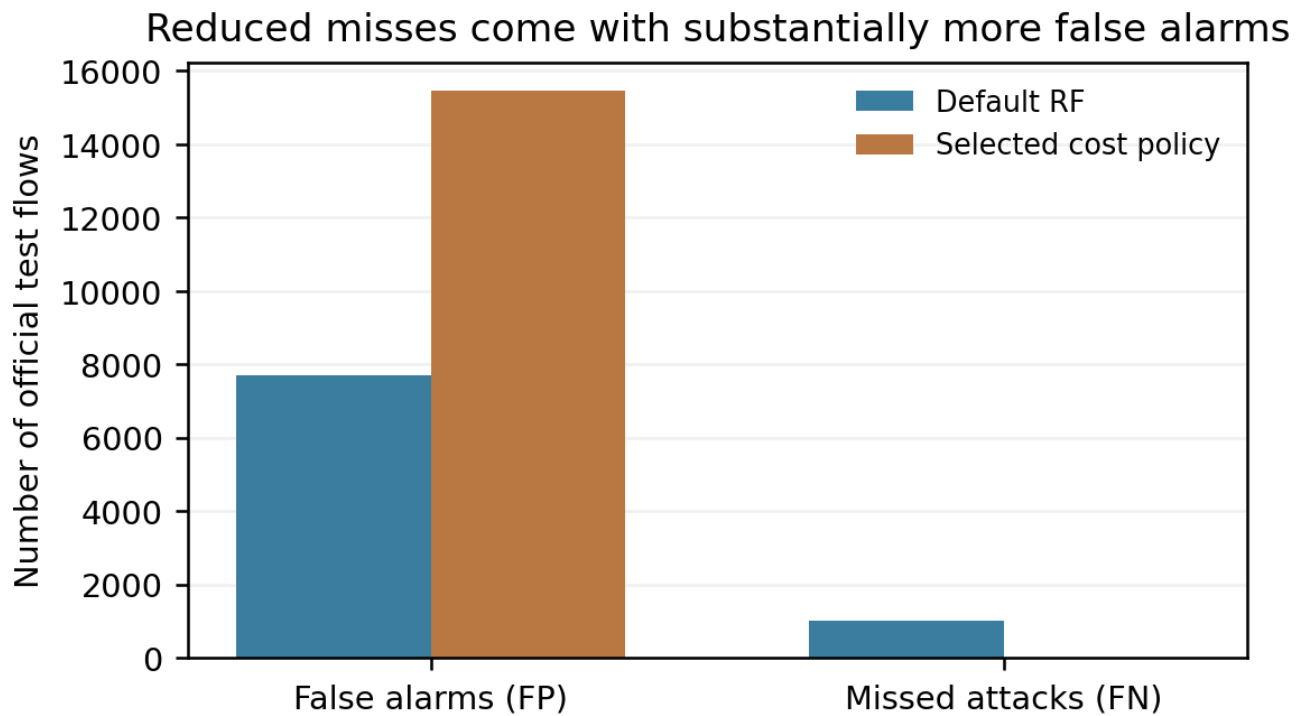


Fig. 4: Absolute test-set false-alarm and missed-attack counts under two operating points on the same random-forest scores.

The prevalence of attack labels declines from 68.06% in official training to 55.06% in official testing (Fig. 2). This class-proportion shift can affect alert precision and reliability but does not, by itself, prove that prior shift caused the threshold-transfer loss; feature or attack-family distributions and record dependencies may also differ. The Brier score improves from 0.0325 (raw) to 0.0308 (calibrated) on the internal selection set, but on the official test partition it deteriorates from 0.0675 to 0.0818. Exploratory ten-equal-width-bin ECE also rises from 0.0751 to 0.0928 on test data. Fig. 5 shows the corresponding binned reliability curves. ECE is binning-sensitive; the result supports a calibration-stability concern, not a universal ranking of calibrators.

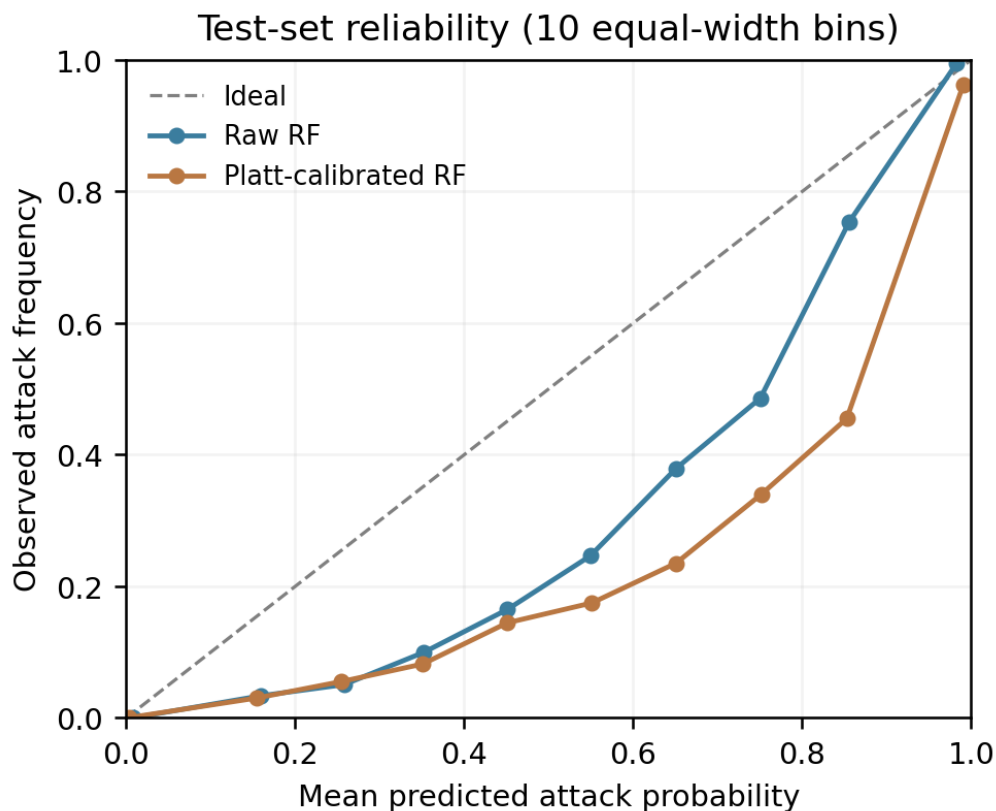


Fig. 5: Test reliability for raw and Platt-calibrated random-forest probabilities. Post-test plot used only for diagnosis; no recalibration on test was performed.

4.3 Random-Seed Sensitivity and Cost Assumptions

To assess dependence on the development partition, the same training/calibration/selection procedure was repeated for three completed random seeds. Table 4 reports the selected threshold and realized costs. All three runs show a validation-set gain that becomes a positive test cost increment. This is a small robustness check, not a broad stochastic or cross-environment guarantee. The official test partition remains the same in all runs.

Table 4: Completed split-seed sensitivity at the illustrative 5:1 error-cost ratio

Seed	Selected threshold	RF cost	Policy cost	Delta cost
42	0.130	0.1548	0.1888	+0.0340
43	0.200	0.1517	0.1826	+0.0309
44	0.185	0.1581	0.1836	+0.0256

The cost ordering is inevitably conditional on the assumed missed-attack/false-alarm ratio. For the fixed seed-42 predictions, the calibrated policy reduces 992 misses at the price of 7,758 extra false alarms. Algebraically, the threshold policy becomes preferable when the ratio C_{FN}/C_{FP} exceeds $7,758/992 \approx 7.82$. Fig. 6 shows a post-hoc cost sensitivity study with the decisions fixed at their originally selected thresholds. At a 10:1 cost ratio, the policy costs 0.1898 per flow against 0.2161 for the default forest. This calculation does not claim that 10:1 represents a validated organizational cost ratio or that the threshold was optimized at 10:1.

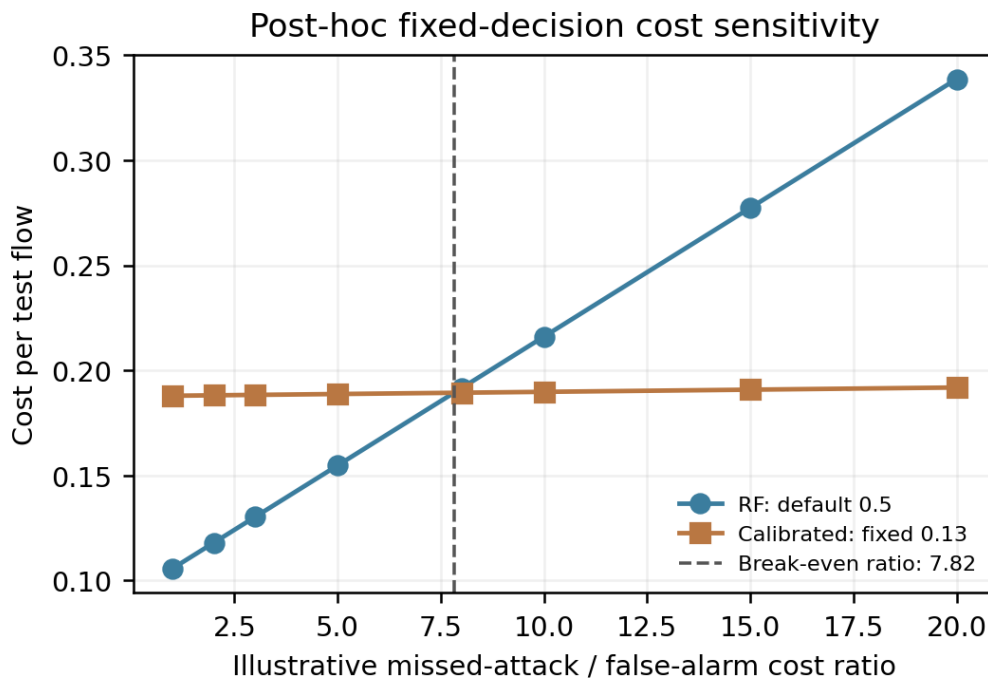


Fig. 6: Cost sensitivity for two fixed test prediction vectors; the dashed line is the algebraic crossover. The figure does not use test labels to choose a new threshold.

4.4 Attack Families and Runtime Context

The seed-42 calibrated policy detects all observed attacks in eight of the nine reported attack families, while its Fuzzers recall is 99.72% (17 of 6,062 Fuzzers observations missed). However, this result must be interpreted alongside the much higher false-alarm count; near-perfect recall from an aggressive threshold is not evidence of detailed attack-family discrimination. The rare Worms class includes only 44 test records, and its observed 100% recall offers limited uncertainty resolution. Moreover, attack_cat labels are not supplied to the binary classifier at inference time.

Measured local batch prediction times in the thesis implementation are approximately 5.0 ms per 1,000 test flows for the forest and 13.2 ms for histogram boosting, excluding collection and preprocessing outside the pipeline. Hardware and runtime conditions were not standardized for deployment analysis, so these measurements are merely within-run engineering context rather than operational throughput claims.

V. DISCUSSION AND LIMITATIONS

5.1 What the Negative Result Means

The same scoring model can produce high ROC AUC and AP while a specific cost-based threshold produces a less desirable outcome. In this experiment, operating-point selection transfers poorly from an internally stratified validation split to the official held-out partition, whose attack share is materially lower. A good internal calibration score does not rule out test-set miscalibration. Conversely, our failure at a 5:1 cost ratio does not imply that cost-sensitive policies always fail; a higher missed-attack cost reverses the ordering for these fixed predictions. Therefore, IDS evaluations should show confusion counts, class prevalence, calibration diagnostics, sensitivity to cost assumptions, and the complete threshold-selection boundary, not merely a maximum-recall headline.

A paired statistical summary strengthens the internal comparison but does not turn this benchmark into a deployment study. The 95% bootstrap interval is conditional on resampling individual records and a fixed test composition. It ignores shared flow sources, time dependencies, host overlap and the sampling process that generated the benchmark. Three development seeds are insufficient for a stable estimate of between-run variability. The threshold grid, model hyperparameters and the baseline family were chosen without an exhaustive pre-registered search; other calibration methods may perform differently. Only the official UNSW-NB15 predefined test split is evaluated, and no attack simulation or live traffic capture was carried out.

The illustrative cost parameters are not validated through incident-response accounting. The category-level observations are descriptive and not a multiclass detector evaluation. The apparent prevalence shift is observed directly; attributing all errors to prior-probability shift would require additional covariate/conditional-shift tests. Finally, historical public datasets can reflect collection artifacts that make generalization to current environments uncertain [4, 3, 15].

5.2 Implications for Future Work

Before recommending deployment, several controlled experiments are needed: temporally or host-disjoint splits, external cross-dataset validation, additional seeds and calibrated uncertainty intervals, explicit measurement of human analyst costs, and threshold revision under detected distribution change. An appropriately designed multiclass study could compare global and per-attack decision rules without using attack-family ground truth at inference [11]. Those experiments should be distinguished from work already completed here.

VI. REPRODUCIBILITY AND RESPONSIBLE USE

The accompanying thesis research project contains preprocessing and analysis scripts, dataset provenance notes, reported measurements, paired comparison outputs, completed seed checks, and figure assets. The official UNSW-NB15 data are obtainable from the dataset providers and are not redistributed with this manuscript package. The seed-42 rerun reproduces all reported forest and calibrated confusion counts exactly. Software was executed with Python 3.13, scikit-learn 1.8.0, pandas 2.2.3 and NumPy 2.3.5 in the recorded environment. All plots are generated from executed scores or documented counts. No private network traffic, personal records or offensive attack-generation code are required.

The publicly accessible UNSW-NB15 dataset authors must receive credit [5, 6]. The project has not undergone independent security deployment evaluation, and results must not be used as a claim that an operational intrusion prevention system is safe. Any independently hosted code and data release should comply with the dataset's terms and institutional policy.

VII. CONCLUSION

A calibration- and cost-selected intrusion-detection threshold can simultaneously improve attack recall and worsen realized operational cost. On the official UNSW-NB15 test partition, lowering the calibrated random-forest threshold from its default decision regime reduced false negatives from 1,009 to 17 while adding 7,758 false positives. Under the explicitly assumed 5:1 cost ratio, test cost increased by 21.95%, despite a large development-set improvement. The paired-bootstrap interval and three completed development-seed repetitions reinforce the direction of this benchmark-specific finding. Reporting only headline recall would reverse its practical interpretation. The principal lesson is methodological: held-out cost transfer, calibration stability and transparent operating-cost assumptions are essential evidence whenever machine-learning security systems are presented as actionable detectors.

DATA AND CODE AVAILABILITY

The official UNSW-NB15 partition files are available from the UNSW Canberra data portal, cited in the references. Analysis scripts and derived results are maintained in the associated academic research project. Public availability has not been independently confirmed. A persistent public repository URL has not yet been established and must not be invented. Authors should confirm that research code and any released derived data are permitted under relevant dataset and institutional terms.

DECLARATION OF COMPETING INTERESTS

Author confirmation required: Each author must disclose relevant financial and non-financial competing interests, or jointly approve an explicit statement that none exist. This placeholder is intentionally not presented as a completed declaration.

FUNDING

Author confirmation required: Record the actual funding source(s), grant number(s), or a verified statement that the study received no specific funding. No funder is presumed.

AUTHOR CONTRIBUTIONS

Author confirmation required: List only authors who meet the journal's authorship criteria and confirm individual contributions (e.g., conceptualization, methodology, software, validation, writing, supervision). No contribution is assigned in advance.

DECLARATION OF AI-ASSISTED TECHNOLOGIES

AI-assisted tools contributed to drafting, language editing and document formatting. Before submission, the authors must review and verify the experimental descriptions, citations, numerical results, intellectual contribution and disclosure text, and adapt this declaration to the journal's current policy. AI tools are not authors.

ACKNOWLEDGMENT

To be completed with confirmed acknowledgements, if any. Do not include unverified institutional assistance or endorsement.

REFERENCES

- [1] K. Scarfone and P. Mell, "Guide to intrusion detection and prevention systems (IDPS)," NIST Special Publication 800-94, 2007, doi: 10.6028/NIST.SP.800-94.
- [2] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, e0118432, 2015, doi: 10.1371/journal.pone.0118432.
- [3] D. Arp et al., "Dos and don'ts of machine learning in computer security," in Proc. 31st USENIX Security Symp., 2022, pp. 3971–3988.
- [4] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," Comput. Security, vol. 86, pp. 147–167, 2019, doi: 10.1016/j.cose.2019.06.005.
- [5] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in Proc. MilCIS, 2015, doi: 10.1109/MilCIS.2015.7348942.

- [6] UNSW Canberra, "The UNSW-NB15 dataset," <https://research.unsw.edu.au/projects/unswnb15-dataset>, accessed Oct. 21, 2026.
- [7] S. M. Kasongo and Y. Sun, "Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset," *J. Big Data*, vol. 7, art. no. 105, 2020, doi: 10.1186/s40537-020-00379-6.
- [8] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. ICML*, 2005, doi: 10.1145/1102351.1102430.
- [9] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017, pp. 1321–1330.
- [10] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950.
- [11] J. Cha, J. Jang, D. Shin, and D. Shin, "Cost-sensitive threshold optimization for network intrusion detection: A per-class approach with XGBoost," *Electronics*, vol. 15, no. 7, art. no. 1542, 2026, doi: 10.3390/electronics15071542.
- [12] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [14] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.
- [15] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, 2014, doi: 10.1145/2523813.