Research Square

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

Label Budget Trade-offs and Calibration Transfer in GNSS Spoofing Detection

bo xie

xiebo@mail.hfut.edu.cn

yi zhao

Research Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11234387/v1

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

Additional Declarations: No competing interests reported.

Label Budget Trade-offs and Calibration Transfer in GNSS Spoofing Detection

Bo Xie[1*] and Yi Zhao[1]

[1]Wuhan Digital Engineering Institute, Wuhan, China.

*Corresponding author(s). E-mail(s): xiebo@mail.hfut.edu.cn;

Abstract

Learning-based global navigation satellite system (GNSS) spoofing detection requires labels both to learn a detection score and to set its alarm threshold. When these information requirements and their acquisition sources are combined, changes in model discrimination can be confused with changes in threshold calibration. We address this evaluation problem through paired experiments that vary calibration size and source while keeping model scores fixed, and compare training–calibration allocations at an equal total label count. The primary SatGrid experiment uses 12,000 correlator observations from three recording cohorts, 135 fitted models and 3,240 threshold conditions. At a budget of 200 labels, a nominal false-positive rate of 5%, and separate-source normal calibration, increasing training support from 5 to 50 observations per class reduces the metric classifier's false-negative rate from 63.60% to 20.30%, while increasing its false-positive rate from 8.83% to 15.13%. The allocation therefore exchanges missed detections for false alarms. Applying the unchanged models and thresholds to 4,000 observations from an additional cohort produces false-negative rates of 99.90% for random forests and 99.95% for the metric classifier at the intermediate allocation. Reversed score ordering distinguishes this failure from threshold miscalibration alone. These results show that an allocation advantage within the primary cohorts can disappear under source transfer. Evaluation should therefore account for both training and calibration labels and verify score ordering and both conditional error rates on separate acquisition sources.

Keywords: GNSS spoofing detection, label budget, threshold calibration, source transfer, false alarms, SatGrid

1 Introduction

Global navigation satellite systems (GNSS) provide positioning and timing information for mobile platforms and networked infrastructure. Spoofing threatens these services by presenting counterfeit satellite signals to a receiver. Receiver-side defenses examine signal power, correlation structure, spatial consistency, or agreement with auxiliary sensors; learning-based approaches use observed signal characteristics to distinguish genuine and counterfeit conditions [1]. Their practical value depends on detecting attacks while controlling false alarms under acquisition conditions that may differ from those used to develop the detector.

Research has progressed along two related directions. Signal-processing methods construct statistics from receiver observables, including frequency-domain correlator behavior [2] and quadrature-channel correlation quality [3]. Learning-based methods represent signal fingerprints and fit decision rules; Spotr demonstrates

a correlator-fingerprinting approach across different collection conditions [4]. More recently, Khan et al. have compared in-domain and cross-domain machine-learning performance on TEXBAT and OAKBAT, showing that favorable mixed-sample results need not persist under dataset transfer [5]. These studies make both the receiver observables and the evaluation partition central to a detector's reported performance.

A further decision lies between fitting a score and deploying an alarm: selecting the threshold. Training observations determine how a model orders samples, whereas trusted normal calibration observations determine which scores trigger an alarm. The area under the receiver operating characteristic curve (AUROC) measures ordering over thresholds; the false-positive rate (FPR) and false-negative rate (FNR) describe a particular operating point. A detector can therefore retain the same AUROC while its observed FPR and FNR change substantially with the calibration data. Reporting training size without calibration size leaves part of the supervised information unspecified.

The separation of score learning and threshold selection is established in Neyman–Pearson (NP) classification [6], including adaptive allocation of samples between these stages [7]. For GNSS receivers, the relevant empirical question also involves the source of the available normal reference data. Additional calibration rows can refine a source-specific threshold, while changes in source can alter both normal and counterfeit score distributions. Comparisons that change fitting and calibration simultaneously cannot determine which stage produced the observed change in errors.

This study examines these interactions on six correlator components from SatGrid [8]. We count the distinct class-labeled training and trusted normal calibration observations under the budget $B = 2k + c$, with k training observations per class and c calibration observations. Three paired comparisons address the research questions: how calibration quantity changes errors with a fixed scorer (RQ1); how reallocating a fixed total count between training and calibration changes the operating point (RQ2); and how these effects differ between logistic regression (LR), random forests (RF), and a metric-learning classifier (RQ3). Calibration-source comparisons are included in RQ1. A separate additional-source evaluation then tests whether the primary observations transfer without updating the model or threshold.

The study makes three empirical contributions. First, it separates fixed-score calibration effects from whole-allocation effects through matched sample identities, nested calibration sets, and explicit source roles. Second, it quantifies the resulting false-alarm–missed-detection tradeoffs across the complete budget matrix, including source-specific outcomes that macro averages conceal. Third, it tests the transfer of these findings using archived models and thresholds, and distinguishes threshold displacement from reversed score ordering through additional-source and saved-score analyses. The primary experiments and the subsequent method extensions are reported separately to preserve their different designs and information access.

The main finding is that a substantial reduction in missed detections at a fixed label count within the primary cohorts does not imply a transferable allocation advantage. Larger calibration sets also change the operating point without repairing an ineffective ordering of the two classes. These observations connect supervision accounting to the actual errors a receiver would incur under the evaluated source conditions. Section 2 places the study in the literature; Sections 3–5 define the problem, data, and methods; Section 6 presents the results; Sections 7 and 8 discuss their implications and conclude.

2 Related work

2.1 Receiver observables and spoofing detection

GNSS interference defenses span signal characteristics, spatial information, and consistency checks, with machine learning offering another means of combining receiver observations [1]. Correlator-based techniques are particularly relevant to this study because they operate on information already associated with signal tracking. Li et al. transform correlator output into the frequency domain, derive a detection threshold under a constant-false-alarm-rate criterion, and evaluate the method using simulation and recorded signals [2]. Mao et al. use early and late

quadrature-channel outputs to construct a signal-quality statistic and evaluate it on TEXBAT and a collected multi-antenna spoofing dataset [3]. Their signal models, temporal processing, and calibration assumptions specify the conditions under which the reported detection rates apply.

Spotr uses physical-layer fingerprints to assess whether GPS signals are consistent with authentic satellite transmissions [4]. SatGrid provides the genuine and spoofing recordings associated with this line of work [8]. We use the released Early, Prompt and Late real and imaginary components as common inputs to three scoring models. This shared representation permits controlled label-budget comparisons. It differs from complete receiver-defense pipelines that also specify feature engineering, temporal aggregation, quality filtering, and alarm logic; published performance values from those pipelines are consequently not interchangeable with the observation-level rates reported here.

2.2 Generalization and evaluation conditions

Khan et al. evaluate four learning methods on TEXBAT and OAKBAT with mixed-sample, pseudorandom-noise (PRN)-related, and cross-domain tests [5]. Their results demonstrate the importance of matching an evaluation partition to the intended generalization claim. A recent preprint likewise reports SatGrid score-order inversion under a power-sweep evaluation (Baweja, 2026, arXiv:2606.22054v2). Its target is predicting degradation from in-distribution score statistics. Our target is the observed FPR/FNR obtained after allocating a specified number of training and normal calibration labels, followed by transfer of the unchanged decision rule.

Evaluation also depends on the decision unit. Zhang et al. study time-aware spoofing evaluation, including detection delay and false-alarm intervals [9]. Such measures require a defined timeline and alarm process. The selected SatGrid observations in our experiment support conditional observation-level error rates. We therefore retain that unit throughout the budget comparisons and additional-source evaluation. This distinction is important when relating sample classification results to an online receiver.

2.3 Calibration and sample allocation

The NP umbrella algorithm formalizes a separation between score estimation and threshold selection, choosing a threshold rank using a type-I error target and a violation probability under the required sampling assumptions [6]. Tong et al. further examine training/threshold sample allocation, including adaptive splitting [7]. These results establish the statistical motivation for counting calibration observations as part of the information used by a classifier.

Here, an empirical normal-score quantile is held constant as the calibration rule while its sample count and source are varied. This design measures the behavior of that rule under the selected acquisition conditions; it does not import the probability guarantees of the NP rank construction. The connection between the two literatures is thus an evaluation question: how do limited supervision and calibration-source variation jointly affect a GNSS detector's operating point? Fixed-score contrasts answer the calibration part, fixed-total-count contrasts answer the allocation part, and unchanged-model transfer determines how far the observed benefits extend.

3 Problem formulation

3.1 Observations and decisions

An observation is $x = (E_R, E_I, P_R, P_I, L_R, L_I)$, where E, P, and L denote the Early, Prompt, and Late correlators. R and I denote real and imaginary components. Each input contains one observation, so the sequence length is $T = 1$. The binary label y identifies Genuine (0) or Counterfeit (1) recording conditions. Attack types and per-observation takeover success are outside this label space.

A scorer $s_m(x; S)$, fitted on support set S, increases toward the Counterfeit class. A normal-only calibration set C determines a threshold tau, giving

$$\widehat{y}(x) = \mathbf{1}\{s_m(x; S) > \tau(C)\}. \tag{1}$$

In Eq. (1), ties are classified as normal. Calibration does not update the encoder, classifier, normalization statistics, or prototypes. Test labels

are used only for evaluation, not to reverse score direction or choose a threshold.

3.2 Source roles and information budget

Let G contain the three source cohorts. Each fold assigns distinct training, separate-calibration, and test sources, denoted a, b, and t. The support set contains k observations per class from a. The near calibration arm selects c normal observations from a reserved pool in a; the heldout arm selects c normal observations from b. Both are disjoint from support and test observations by sample identity. Near calibration deliberately shares an acquisition source with training and is not an independent recording.

Each test source supplies all 4,000 observations in its selected pool, with 2,000 per class. No test-source observations enter the fold's preprocessing or fitting. All three rotate through primary evaluation. A subsequently selected Blacksburg cohort is reserved for the additional confirmation described in Section 4.5; its model performance is not used to modify the primary experiment.

Using unique support and calibration identities, the per-condition information budget is

$$B = |I_S \cup I_C| = 2k + c. \tag{2}$$

Equation (2) counts prototype construction as support reuse, without additional labels. Unselected candidates are excluded; default thresholds use 2k labels. Conditions overlap, so their budgets cannot be summed as disjoint collections. Trusted normal calibration and class labels are assumed supplied, without equal acquisition costs.

3.3 Comparison targets and aggregation

For RQ1, matched calibration conditions fix the model, test identities, and test scores. Size comparisons use nested normal samples; source comparisons use equal c. For RQ2, allocations (5,190), (20,160), and (50,100) all satisfy $B = 200$. Both model training and calibration change, so their separate causal contributions are not identified by this contrast. RQ3 compares methods at matching support identities, role, seed, k, c, and nominal level.

We report FPR=FP/(FP+TN), FNR=FN/(FN+TP), and attack-oriented AUROC. Both conditional denominators contain 2,000 observations per test source. The balanced evaluation pool does not represent deployment attack prevalence or establish alarm precision.

For an error rate e, we first average the two roles and five seeds within each test source, then average the three sources equally:

$$\bar{e} = \frac{1}{3} \sum_{t \in G} \left[\frac{1}{10} \sum_{j=1}^{2} \sum_{r=0}^{4} e_{t,j,r} \right]. \tag{3}$$

The weights in Eq. (3) do not create independent replications. Reused models, observations and seeds preclude population inference here. Source-specific rates accompany macro means; worst-source FPR and FNR can occur on different sources.

4 Data and controlled protocol

4.1 Dataset coverage

The pool uses SatGrid version 2 [8], with eight MAT files from three Arlington cohorts. Two Nov8 rounds are conservatively combined to avoid treating potentially shared acquisition circumstances as independent train/test sources. Table 1 reports coverage. Genuine and Counterfeit files have matching PRN quotas within each recording pair. Across cohorts, PRNs, environment, directory levels, and processing context vary together; cross-source differences cannot be attributed solely to date, satellite, or physical spoofing power.

Counterfeit/0 remains labeled Counterfeit. Labels follow release conditions and are never reassigned from feature amplitude or predictions. Feature order is fixed, with shape $[N, 1, 6]$. Explicit mappings reconcile source field names, and available equivalent fields are checked for equality. PRN, filename, cohort, source-row index, directory level, carrier-to-noise measurements, and lock metadata are excluded from model inputs. Consistent field mapping does not establish identical receiver settings or physical calibration across files.

4.2 Selection and quality handling

Each file's source-row range is divided into a predetermined number of equal-index-length strata.

Table 1 Selected file coverage. The Counterfeit directory number is a source identifier, not a calibrated physical power measurement

Source cohort	Archive file ID	PRN	Genuine rows	Counterfeit rows	Counterfeit directory level	Files
Aug23	27236452	2	2000	2000	0	2
Sep10	27238418	5	2000	2000	0	2
Nov8, round 1	27237007	2	1000	1000	30	2
Nov8, round 2	27237459	10	1000	1000	30	2
Total	Four archives, three cohorts	—	6000	6000	—	8

One row per stratum is selected by a fixed hash rule based on source-file identity and stratum index. Model scores and class separability do not influence selection. File and source-row identities are preserved, and equal class/cohort quotas prevent file length from determining aggregate weights.

All 20 selected observations whose recorded carrier-to-noise field (CN0) is zero are retained: 12 Genuine and eight Counterfeit. The selected six-component vectors are finite, their sample identities are unique, and no exact six-component duplicates were found in the selected pool. These checks do not rule out approximate duplicates or upstream waveform reuse. Sparse index sampling is neither uniform physical-time sampling nor proof of independence. No sliding windows or inferred timestamps are introduced.

4.3 Role permutations and calibration pools

Figure 1 summarizes all six assignments of training, separate-calibration, and test sources. Each training source reserves 500 normal candidates for near calibration, excluding those rows from its fitting candidates. Heldout calibration draws from the separate source's 2,000 normal candidates. Only the selected c scores reach threshold estimation; unused candidates and the separate source's Counterfeit observations do not contribute to fitting or calibration.

Support and calibration draws are stratified by archive and per-row PRN. Within strata, sample identities and the seed define a fixed hash order. Samples are taken cyclically across strata; increasing k or c takes a longer prefix of the same ordering. Seeds are 0–4, and methods share selected identities under matched settings. Candidate counts and actual observed counts are recorded separately.

4.4 Frozen experimental matrix

Table 2 distinguishes two complementary designs. In fixed-training comparisons, only calibration information changes. In fixed-total-budget comparisons, both support size and calibration size change. The primary nominal level is 5%, with 1% as a sensitivity level. The complete grid was fixed before model-performance inspection on this pool, following data-format and coverage diagnostics. The later additional-source and model-extension phases are distinguished below.

Caching by training source, method, seed and k gives 135 models and 270 model-test blocks. The 3,240 cells are threshold evaluations, not independent acquisitions. AUROC counts unique blocks. Contrasts use near calibration, $c = 20$ and allocation (20,160) as the respective references.

4.5 Additional source confirmation

After inspecting the primary results, we selected the SatGrid Blacksburg Sep28 G4/S4 pair, with PRN 9 verified across both full files. Selection used documented source identity, native field compatibility and the smallest common directory PRN, not model performance. Other files in this cohort had already undergone format/content auditing. Using the original deterministic hash-jittered source-index stratification, we froze 2,000 rows per condition before scoring. All selected rows, including 12 Genuine and 2 Counterfeit CN0-zero rows, remain in the primary confirmation.

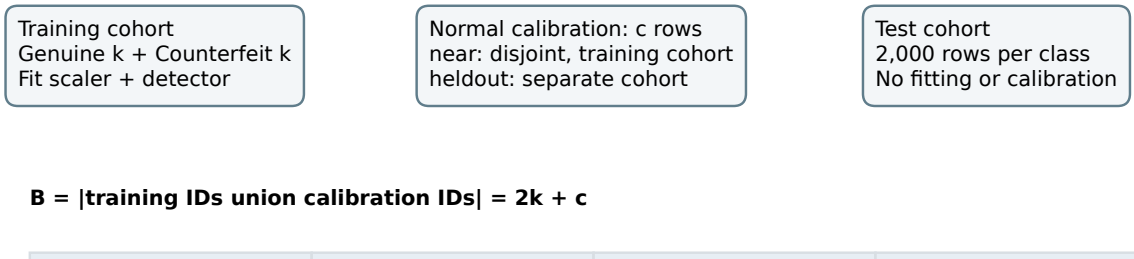


Fig. 1 Source roles and visible supervision. Near calibration uses normal rows from the training source that are disjoint from support. Heldout calibration uses a third source. A model fitted on one training source is shared across its two test roles. Six role assignments represent only three distinct test cohorts

Table 2 Complete condition matrix. Each row includes three methods, six roles, five seeds, two calibration arms, and two nominal levels

Design	Training k per class	Normal calibration c	Budget B	Evaluation cells
Fixed training	5, 20	20, 100, 500	2k+c	2160
Fixed total budget	5, 20, 50	190, 160, 100, respectively	200	1080
Total	Nine distinct (k,c) settings	near and heldout	—	3240

All 135 archived models and 3,240 thresholds are reused. No new-source observations enter fitting, scaling or calibration. All six original training/calibration role assignments and five seeds are retained; their equal-weight mean describes one new test cohort, not 30 independent acquisitions. Near-arm repetitions across held-out-role choices remain explicit. Fixed scores and thresholds also support a prespecified sensitivity excluding CN0-zero test rows only.

Scores and thresholds retain the original float64 comparison convention. An independent reconstruction recomputes the saved scores, thresholds and error counts; the implementation correction and validation details are documented in Online Resource 3. The additional cohort changes several acquisition factors jointly, providing a transfer test across collection conditions within the same dataset.

4.6 Additional methods and exploratory development

Subsequent studies used all four already exposed cohorts. First, a fixed XGBoost [10] extension reused the primary support/calibration identities and both test domains, adding 45 estimators and 2,160 calibrated cells. Second, a distinct four-source leave-one-source-out development design

pooled support across the other three sources: 20 observations per class in total, plus 160 disjoint normal calibration rows. Seeds 0 and 1 and all 4,000 validation-source rows were retained. This protocol differs from the primary one-source training design; its rates are not interchangeable with Table 6.

Three sequential development rounds evaluated 80, 72 and 32 models, including paired controls. Later rounds used earlier results, so these analyses are reported as adaptive development rather than additional confirmations. The four evaluated cohorts contain 16,000 labeled observations. Thus $B = 200$ is a per-condition training/calibration count; the full development process accessed a larger labeled research pool. Within each fold, the target source is excluded from preprocessing, training and calibration. Online Resource 5 reports every tested configuration and its source-specific results.

5 Scorers and normal threshold calibration

5.1 Common preprocessing and conventional baselines

LR, RF and the metric-learning classifier provide linear, tree-ensemble and learned-embedding scoring functions, respectively (Table 3). Their common inputs, selected labels and calibration rule allow the budget effects to be compared across distinct model families. All receive identical six-component support observations. StandardScaler is fitted only on the current 2k support rows and is reused unchanged for calibration and testing. No additional pretraining or test-guided hyperparameter search is performed. With $T = 1$, the conventional methods' temporal aggregation reduces to the single observation.

LR and RF use scikit-learn [11]. LR is LogisticRegression with lbfgs, C=1, tolerance 10^{-4}, and at most 3,000 iterations. RF [12] contains 300 trees with Gini splitting, square-root feature subsampling, bootstrap sampling, no maximum-depth limit, and minimum leaf size one. Its random state follows the experimental seed. Both classifiers are fitted once per support set, without refitting during calibration. The actual Counterfeit class column supplies their attack probabilities; maximum class confidence is not substituted for that directed score.

5.2 Metric-learning scorer

The metric implementation uses a single bidirectional LSTM layer [13], followed by pooling, dropout 0.2, a 32-dimensional projection, tanh, and L2 normalization. The embedding feeds a classification head during training. The objective combines cross-entropy with weight one and batch-hard triplet loss (Hermans et al., In Defense of the Triplet Loss for Person Re-Identification, 2017, preprint arXiv:1703.07737v4) with margin 0.3. Adam uses learning rate 0.001 and zero weight decay. Each of 40 epochs uses the complete support set as one batch; no class-balanced batch sampler or adaptive margin is used.

After training, evaluation-mode support embeddings are averaged within each class to obtain prototypes p_0 and p_1. The attack score is

$$s_{\mathrm{metric}}(x) = \|z(x) - p_0\|_2 - \|z(x) - p_1\|_2. \quad (4)$$

Prototypes are not renormalized, and their construction neither updates the encoder nor consumes additional labels. Calibration and testing use Eq. (4), not classification-head probabilities. Each input has $T = 1$, so this configuration learns an observation embedding rather than exploiting inter-observation temporal context.

The exploratory extension also evaluates TabM [14] and ModernNCA [15] using their author implementations. Both operate on the same numerical inputs and budget within the four-source development protocol. Their configurations and preprocessing variants are reported in Online Resource 5.

5.3 Empirical rank threshold

Let $s_{(1)} \le \cdots \le s_{(c)}$ be the sorted normal calibration scores. For a nominal level alpha, the frozen rule is

$$r = \lceil (1-\alpha)c \rceil, \qquad \tau = s_{(r)}, \qquad \widehat{y} = \mathbf{1}[s(x) > \tau]. \quad (5)$$

Table 3 Scorers and frozen settings

Method	Model	Key settings	Attack score
LR	Logistic regression	lbfgs, C=1, max_iter=3000, tol=10^{-4}	Counterfeit probability
RF	Random forest	300 trees, Gini, sqrt, bootstrap	Counterfeit probability
metric	Bidirectional LSTM and class prototypes	hidden=64 per direction, proj=32, 40 epochs, cross-entropy plus batch-hard triplet loss, margin=0.3	Distance to normal prototype minus distance to Counterfeit prototype

In Eq. (5), alpha is an integer fraction and the rank uses integer arithmetic. Scores are neither interpolated nor jittered; ties remain normal. For $c = 20$ and alpha=0.01, the threshold is the calibration maximum. Zero empirical calibration errors in that case do not establish accurate estimation of a 1% population tail.

The nominal level specifies the empirical quantile; observed error rates on a different source are measured separately. The relation to NP threshold selection [6] is discussed in Section 2.3. Default references use strict thresholds of 0.5 for LR/RF and zero for metric, at budget 2k. Test AUROC never determines score direction.

5.4 Implementation and reproducibility

All experiments run on CPU with one numerical-computation thread. The recorded environment comprises Python 3.14.5, NumPy 2.5.3, scikit-learn 1.9.0 and PyTorch 2.11.0+cu128. Effective estimator parameters, selected sample identities, fitted scalers, class mappings, prototypes and model states are preserved. An independent validation path reconstructs scores, thresholds, budgets and error counts and checks completeness of the experimental matrix. The supplementary resources document the experiment settings and complete results.

Generative AI tools assisted software development, debugging, experiment checking, literature retrieval and manuscript drafting.

6 Results

6.1 Ranking and default operating points

Figure 2 and Table 4 separate ranking from threshold choice. LR's source-macro AUROC remains approximately 0.498 across the three training budgets. RF rises from 0.854 at $k = 5$ to 0.903 at $k = 20$, then gives 0.897 at $k = 50$. Metric increases from 0.779 to 0.934. Source-specific results below show how these macro values vary across acquisition cohorts.

At $k = 20$, RF's source-specific AUROCs are approximately 0.987 for Aug23, 0.724 for Nov8, and 0.999 for Sep10; metric gives 0.931, 0.786, and 0.963, respectively. The macro averages therefore conceal substantially weaker ranking on Nov8. Changes in PRN, acquisition background, and other source factors are confounded, so the difference does not isolate a date or power effect.

6.2 Nominal levels and observed errors

Table 5 presents all 12 method/arm/level combinations at the middle allocation $(k, c) = (20, 160)$, $B = 200$. This allocation is the middle point of the fixed-budget design. Full results contain 108 source-macro and 324 source-specific rows.

All 12 macro FPRs exceed their respective nominal levels. This does not imply an exceedance for every source or seed. Under heldout calibration at 5%, metric lowers FPR relative to RF by about 1.73 percentage points but raises FNR by about 12.39 points. LR attains a smaller FPR of 7.80% with FNR 99.82%; across its 36 macro conditions, FNR ranges from 98.9333% to 99.9950%. Reporting its lower FPR without the corresponding missed-detection rate would conceal its failure to detect almost all Counterfeit observations under these calibrated settings.

6.3 Calibration quantity with fixed training

Figure 3 shows the complete primary-level calibration-size comparison, with the 1% counterpart in Supplementary Fig. S1 (Online Resource 1). Models and test scores are fixed within each

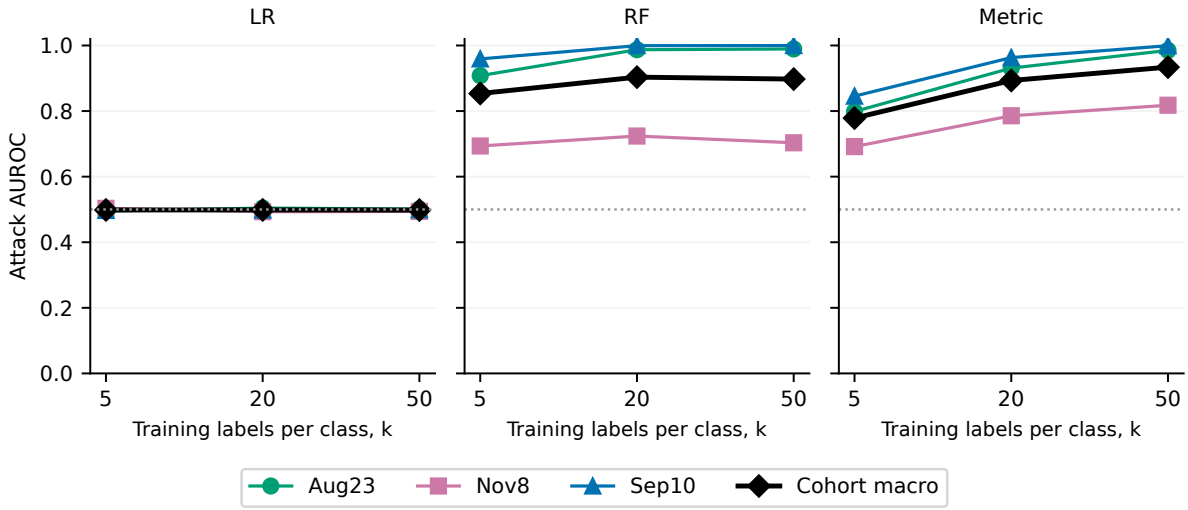


Fig. 2 Ranking across training budgets. Source curves average roles and seeds; the black curve is the equal-weight source mean. Each model–test block is counted once, independently of calibration-cell multiplicity

Table 4 Source-macro AUROC and default-threshold errors. Percentages refer to conditional observation-level rates. Default thresholds use no calibration labels

Method	k	Default budget	AUROC	FPR (%)	FNR (%)
LR	5	10	0.498421	61.24	9.71
LR	20	40	0.498571	58.79	12.36
LR	50	100	0.498125	57.80	14.17
RF	5	10	0.853536	45.38	12.69
RF	20	40	0.903449	20.72	23.51
RF	50	100	0.897492	15.77	26.36
metric	5	10	0.778583	54.45	9.23
metric	20	40	0.893271	46.53	5.91
metric	50	100	0.934000	41.81	4.92

contrast, so changes in FPR/FNR occur without changing AUROC.

For $k = 20$ and heldout calibration, increasing c from 20 to 500 changes metric's FPR from 13.88% to 10.60% and FNR from 33.79% to 39.13%. RF changes from 16.68%/21.25% to 12.42%/26.03%. In both specific contrasts, all three source means exhibit lower FPR and higher FNR. Additional normal calibration information changes the operating point, but the FPR reduction does not constitute improvement in both errors. At $c = 500$, both macro FPRs still exceed 5%. The other allocations and the 1% sensitivity condition are included in the complete comparison.

6.4 Calibration-source dependence

Figure 4 displays heldout-minus-near contrasts across all nine (k,c) settings. Each comparison changes only normal calibration source. Negative differences indicate reductions in the corresponding error, and both macro and individual-source points are retained.

At $k = 20$, $c = 160$, and the 5% level, heldout calibration reduces RF's macro FPR by 5.19 points but increases FNR by 5.65 points. FPR decreases in all three sources, while FNR increases by 11.67 points on Nov8 and 5.43 points on Sep10. For metric, macro FPR and FNR both decrease, by 1.41 and 2.29 points. Nevertheless, Nov8 FPR increases by 4.47 points and Sep10

Table 5 Calibrated source-macro errors at $k = 20$ and $c = 160$

Method	Calibration arm	Nominal FPR (%)	Test FPR (%)	Test FNR (%)
LR	near	5	8.77	99.73
RF	near	5	17.14	20.84
metric	near	5	11.62	41.15
LR	heldout	5	7.80	99.82
RF	heldout	5	11.95	26.48
metric	heldout	5	10.21	38.87
LR	near	1	3.41	99.97
RF	near	1	7.30	29.24
metric	near	1	3.59	55.77
LR	heldout	1	2.96	99.99
RF	heldout	1	4.65	37.72
metric	heldout	1	4.60	52.33

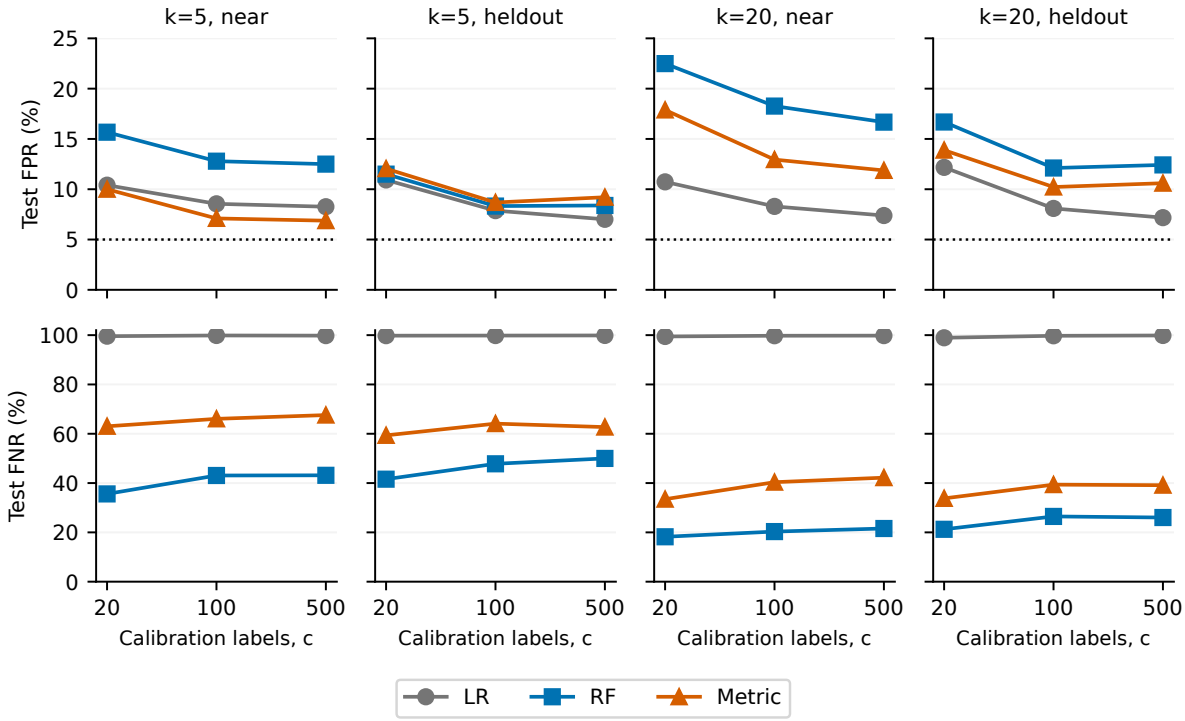


Fig. 3 Calibration-size comparisons at the nominal 5% level. Columns separate training budget and calibration arm. The top and bottom rows show FPR and FNR on different scales; each rate uses a common scale across columns. Dashed lines denote the nominal level

FNR increases by about 0.85 points. A favorable average therefore does not establish uniform benefit.

Joint macro error reductions combine different folds; moving one threshold on fixed scores cannot reduce both errors. Near and heldout name source roles, not measured distribution distances. The measured source-specific changes determine the benefit of each calibration arm.

6.5 Allocation under a fixed supervision count

Figure 5 compares the three $B = 200$ allocations. Unlike the preceding calibration contrasts, both

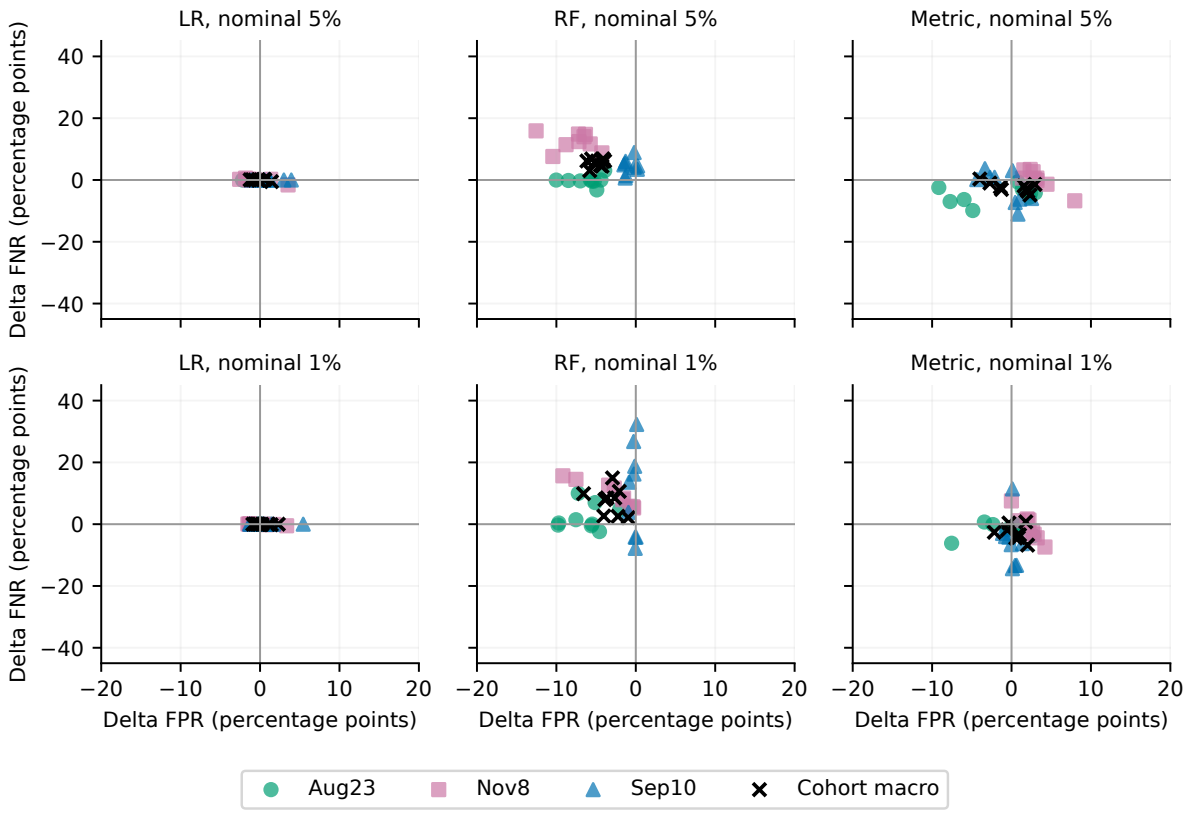


Fig. 4 Paired calibration-source changes. Axes are percentage-point differences in FPR and FNR. Colored points are source means and black crosses are source-macro means, covering both nominal levels and all allocations. Overlap can conceal coincident points; exact values are retained in the figure data. Point counts are not counts of independent acquisitions

scoring models and calibration observations can change. The comparison evaluates whole allocations rather than isolating the effect of increasing training size alone.

Under heldout calibration at 5%, moving metric from (5,190) to (50,100) lowers FNR from 63.60% to 20.30%, a reduction of about 43.30 percentage points. FPR rises from 8.83% to 15.13%, about 6.30 points. Without a specified utility or acceptable-error constraint, the allocation with the lowest FNR cannot be called the optimal deployment policy.

RF also resists a simple more-training-is-better interpretation. Under the same conditions, its (20,160) FPR/FNR is 11.95%/26.48%, versus 14.38%/26.97% at (50,100). Both macro errors are slightly higher in the latter condition. LR retains very high FNR across these allocations. The allocation preference therefore depends on both the scoring model and the acceptable balance between false alarms and missed detections.

6.6 Local failures behind macro averages

Figure 6 expands the middle allocation by test source. For RF with heldout calibration at 5%, Aug23 FPR is approximately 25.34%, while Nov8 FNR is approximately 73.00%. These different sources account for the worst values behind the 11.95%/26.48% macro point. Metric's corresponding worst-source FPR and FNR are approximately 17.65% and 66.02%.

Equal supervision counts permit different error trade-offs, while macro summaries can obscure local failures. A candidate operating point should consequently be assessed against each source as well as the macro mean.

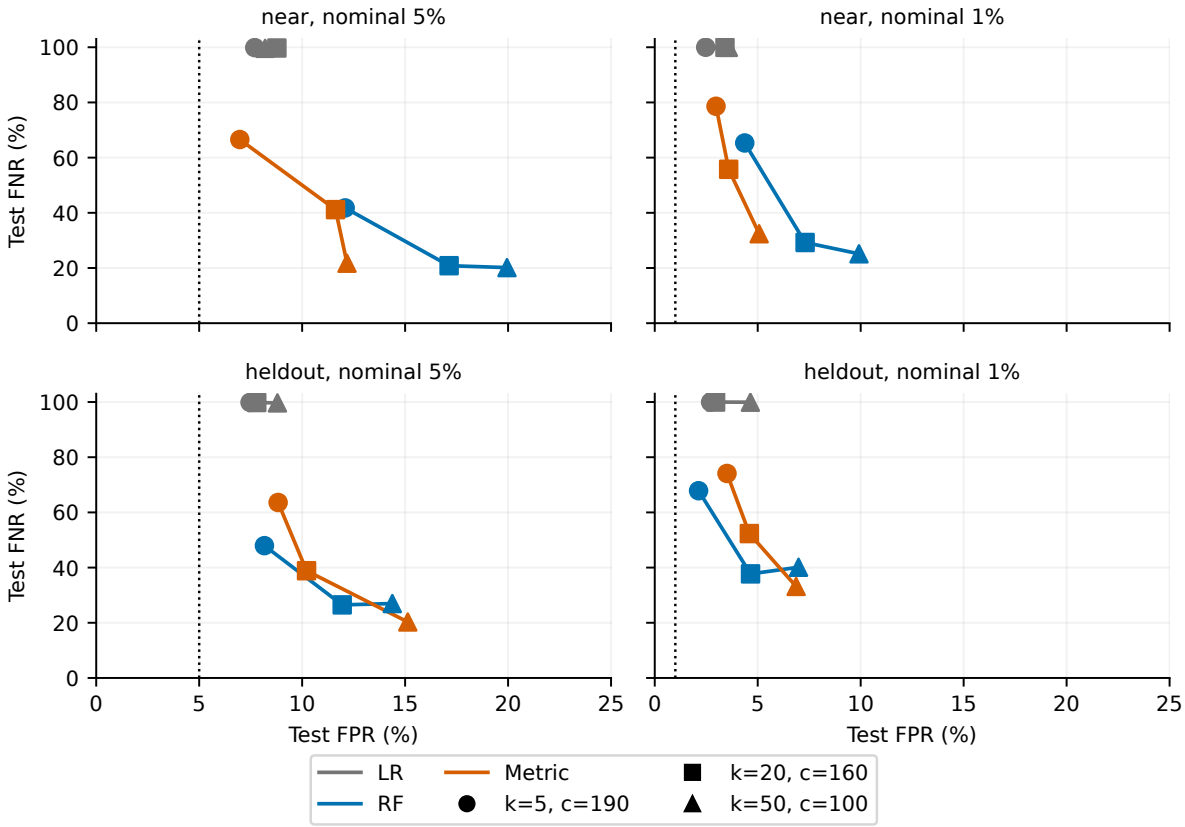


Fig. 5 FPR–FNR operating points at $B = 200$. All three methods, allocations, calibration arms, and nominal levels are shown, giving 36 macro conditions. Lines connect the specified allocations; they are not trajectories learned by a budget optimizer

6.7 Confirmation under changed acquisition conditions

Table 6 compares exactly matched old and new test conditions. RF and metric nearly always miss Counterfeit observations on the added cohort. Their attack-oriented AUROCs are far below 0.5, indicating reversed ordering under the fixed direction rather than only an unfavorable threshold. Score direction is not reversed after observing labels. LR has smaller calibrated errors than on the old sources, but its mean AUROC remains approximately 0.504.

At $B = 200$, changing (k,c) from (5,190) to (50,100) increases metric FPR from 25.125% to 45.052%, while FNR changes only from 99.968% to 99.915%. RF FPR increases from 22.858% to 34.437%, with FNR increasing from 99.673% to 99.905%. The earlier substantial miss-rate benefit does not transfer. At fixed $k = 20$, $c = 20 \rightarrow 500$ reduces RF and metric FPR by 11.905 and 6.158 points, but neither scorer recovers useful attack detection. Calibration still changes operating points while ranking remains ineffective.

Across all conditions, excluding only CN0-zero test rows changes FPR/FNR by at most 0.244/0.064 points. This limited check cannot explain the large degradation or rule out other acquisition effects. The confirmation supplement reports all 108 condition means, 648 role means and 2,700 matched contrasts, including the nominal 1% sensitivity conditions.

6.8 Post-confirmation method coverage and development

At the intermediate primary allocation, XGBoost produces FPR/FNR of 8.323%/39.840% on Arlington and 18.722%/99.533% on Blacksburg; AUROC changes from 0.847818 to 0.214309. This

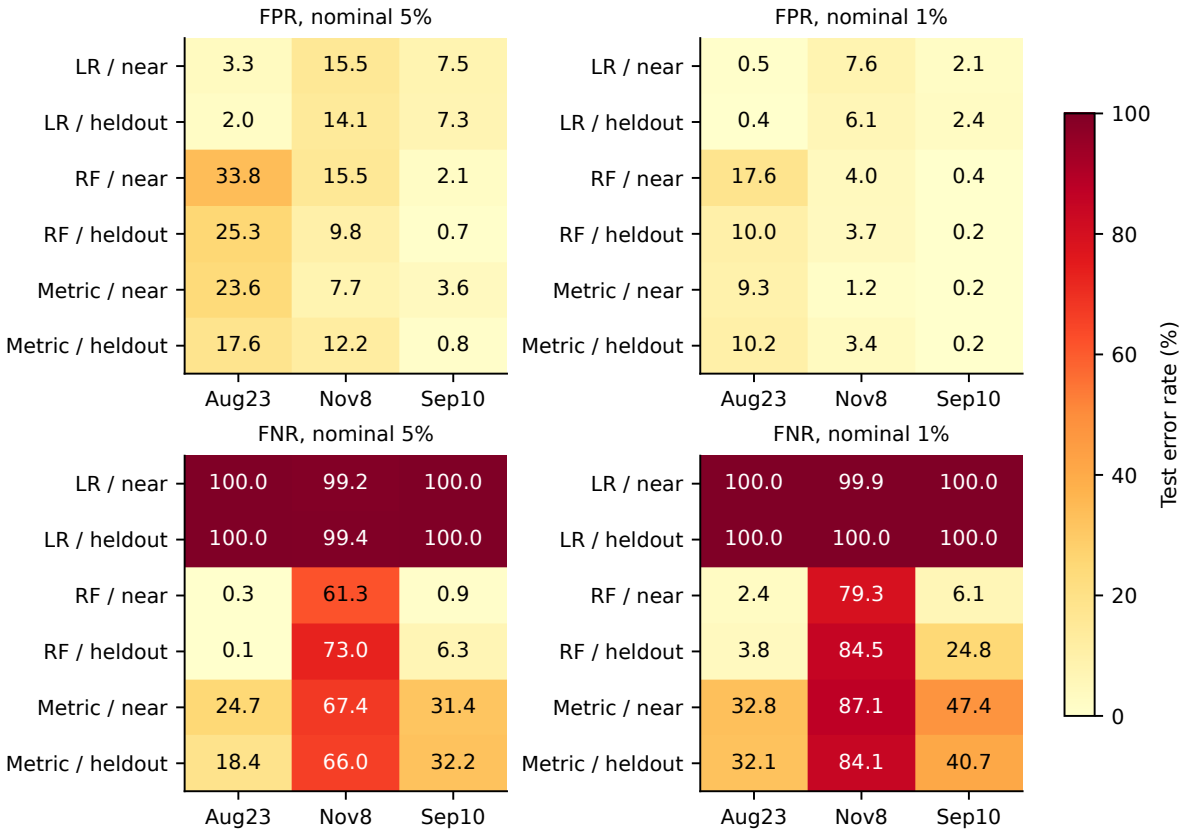


Fig. 6 Source-specific errors at $k = 20$ and $c = 160$. Each value averages two roles and five seeds. Panels share a 0–100% color scale; labels use one decimal place. Exact tables support comparisons of small differences. The figure supplements rather than replaces the full budget grid

Table 6 Intermediate allocation $k = 20$, $c = 160$, heldout calibration and nominal 5% level. New values average six source roles and five seeds; old values use paired original cells. Differences combine several acquisition changes and are not causal effects

Method	Old FPR (%)	New FPR (%)	Old FNR (%)	New FNR (%)	New AUROC
LR	7.795	0.220	99.815	53.532	0.503782
RF	11.945	32.728	26.482	99.902	0.026630
metric	10.213	33.695	38.867	99.947	0.014780

extension was specified after the additional-source results had been observed; its full matrix is reported in Online Resource 4.

In the different four-source development protocol, a support-neighborhood candidate with signed amplitude compression has mean FPR/FNR of 6.681%/70.463%; TabM and ModernNCA with the same compression give 7.313%/61.488% and 5.438%/70.406%. Blocking same-source neighbors during candidate training changes FPR/FNR to 6.756%/66.438%, but Blacksburg FNR remains 99.875%, and the worst-source balanced error increases. Raw-input source masking worsens mean FNR. The remaining high Blacksburg miss rate persists across these development changes. Full method and source results, including unsuccessful configurations, are retained in Online Resource 5.

7 Discussion

7.1 Separating threshold effects from score-order failure

The controlled comparisons distinguish two mechanisms. In the primary experiment, calibration quantity and source move the operating point while leaving the model and its test-score ordering unchanged. For example, additional heldout calibration observations reduce RF false alarms but increase missed detections. This is a threshold trade-off: the score retains discrimination, but a more conservative alarm decision excludes some counterfeit observations.

The Blacksburg results show a different pattern. RF and metric AUROCs of 0.026630 and 0.014780 indicate that the fixed attack-oriented scores order the two classes largely in the wrong direction. Recalibrating a scalar threshold on that score can change the number of alarms, but cannot change the pairwise ordering measured by AUROC. The large additional-source FNR therefore cannot be interpreted solely as an inaccurate choice of the normal quantile. The score representation and its relationship to the class conditions also require examination.

LR provides a third diagnostic reference: scores with little class separation. Saved-score analysis across 90 blocks gives AUROCs from 0.469366 to 0.541400. On average, 97.11% of Counterfeit scores lie inside the Genuine interquartile interval, and the median ratio of Counterfeit to Genuine interquartile ranges is 0.0745. In 1,031 of 1,080 calibrated LR cells, the threshold is at least the Counterfeit 99th percentile. Normal-tail thresholding consequently removes almost all Counterfeit alarms along with many Genuine alarms. Supplementary Figs. S2 and S3 (Online Resource 2) show the score distributions and the distinction between calibration and test FPR.

These patterns explain why either ranking or one error rate alone is insufficient. Good primary-source ranking may not transfer; a reduced FPR may accompany near-complete missed detection; and a threshold change can alter both errors without any change in representation quality.

7.2 What the label budget comparison resolves

Counting $B = 2k + c$ makes the reference information used for threshold selection visible alongside class-labeled training data. At $B = 200$, the metric classifier's (50,100) allocation has a much smaller primary-source FNR than (5,190), but incurs more false alarms. RF exhibits a different preference: at the same nominal level and calibration arm, the middle allocation has slightly smaller macro errors than the larger-training allocation. Thus a training-size comparison made without its calibration count would describe a different decision problem.

The appropriate allocation also depends on the intended source conditions. The primary metric FNR reduction of 43.30 percentage points becomes only 0.053 percentage points on Blacksburg, while its FPR increases substantially. More training labels from the existing source do not recover the primary benefit in this transfer test. A fixed total count therefore makes allocations comparable, but the selected sources determine the relevance of that comparison to a deployment.

For evaluation practice, these findings support reporting three items together: the disjoint observations consumed by fitting and calibration; the calibration-source relationship to training and testing; and source-specific FPR/FNR alongside AUROC. An application-specific choice among the measured operating points additionally requires an acceptable false-alarm or missed-detection constraint. The complete grid provides the corresponding trade-offs without selecting a policy retrospectively from test outcomes.

7.3 Source-dependent observables

A post-hoc source-file check re-read the selected rows from all ten MAT files. All 96,000 component values match the combined 16,000-row pool, and labels agree with the release directories. This check finds no import-column or condition-label reversal in the analyzed rows.

The raw six-component amplitude exhibits an opposite class association across the two collection settings. Its L2-amplitude medians (Genuine/Counterfeit) are 52,223/1,691 for Aug23, 54,056/18,880 for Nov8, 89,274/2,071 for Sep10, and 17,468/1,883,315 for Blacksburg. Taking

larger amplitude as the Counterfeit direction gives descriptive univariate AUROCs of 0.000034, 0.139750, 0.000000 and 0.999500, respectively. This observation is consistent with a vulnerability to source-dependent class cues, but does not identify the features used internally by each fitted model. Physical scaling and integration comparability across files remain unverified. The diagnosis leaves the original labels, model directions and thresholds unchanged (Supplementary Fig. S4, Online Resource 5).

Zero-CN0 observations have little effect in a fixed-model sensitivity check. Removing only these test rows changes FPR/FNR by at most 0.173332/0.136774 percentage points in the primary matrix and 0.244/0.064 points in the additional-source matrix. Those rows alone therefore do not account for the much larger transfer degradation. Other quality effects and quality-filtered retraining require separate investigation.

7.4 Scope and limitations

The evidence covers three Arlington cohorts and one Blacksburg cohort from SatGrid. Dates, PRNs, environments and processing context vary together, and the design does not isolate their causal contributions. Grouping the two Nov8 rounds reduces the risk of splitting related collections across source roles. Nevertheless, sparse source-index sampling does not establish independent observations or uniform physical-time sampling. We therefore use descriptive source-level summaries rather than population significance claims based on repeated seeds or role permutations.

The task is observation-level classification of Genuine and Counterfeit recording conditions. The labels do not specify attack subtype or receiver takeover, and the selected rows do not support event latency or false alarms per hour. Balanced test pools estimate the two conditional error rates but not deployment alarm precision. Trusted normal calibration data are assumed to be available; their acquisition and authentication costs can differ from those of class-labeled training observations.

The shared six-component input isolates budget comparisons across scorers. It omits metadata and temporal sequences, so the recurrent metric implementation receives $T = 1$. Complete GNSS defense pipelines with additional observables or temporal decision logic address a different input setting. The later TabM, ModernNCA and candidate comparisons use a separate, adaptively developed four-source protocol with fixed training settings. They extend the diagnostic evidence rather than provide a tuned ranking of all available detectors. All four evaluated cohorts have informed the research process; further method development requires additional held-out acquisition conditions for an independent assessment.

8 Conclusion

This study quantified how training labels, normal calibration labels and their acquisition sources jointly affect GNSS spoofing detection. Paired fixed-score comparisons separated threshold effects from model changes, and fixed-total-count comparisons exposed the false-alarm–missed-detection trade-off of reallocating supervision. Within the primary SatGrid cohorts, increasing the metric classifier's training share at $B = 200$ reduced FNR by 43.30 percentage points while increasing FPR by 6.30 points. The unchanged models and thresholds then failed to retain that benefit on an additional cohort, where RF and metric missed more than 99.9% of Counterfeit observations at the intermediate allocation.

The combination of budget comparisons and source transfer establishes a practical evaluation requirement: allocation results must be interpreted jointly with calibration-source assumptions and the stability of score ordering. A lower error on the development sources is insufficient evidence for selecting a transferable operating point. Complete supervision accounting and separate-source verification make these dependencies observable and identify whether the next intervention should address the threshold, the score representation, or acquisition-dependent inputs.

Statements and Declarations

Funding. The authors received no funding for this work.

Competing interests. The authors have no relevant financial or non-financial interests to disclose. No patent application related to this work has been filed.

Author contributions. Bo Xie contributed to research design, software implementation, manuscript drafting and revision, and project supervision. Yi Zhao conducted the experiments and data analysis.

Data availability

The original SatGrid recordings are available from the dataset publisher [8]. The supplementary resources provide the selected-source description, experiment settings and complete result tables. Code and trained model artifacts have not been deposited in a public repository.

Supplementary information

Online Resource 1 (ESM_1.zip): complete primary source-macro and source-specific results, with the 1% calibration-size sensitivity plot (Fig. S1).

Online Resource 2 (ESM_2.zip): saved-score and threshold diagnostics, including LR score distributions and calibration/test false-positive rates (Figs. S2–S3).

Online Resource 3 (ESM_3.zip): additional-source evaluation, numerical implementation amendment, 108 condition means, 648 role means and 2,700 paired contrasts.

Online Resource 4 (ESM_4.zip): post-confirmation XGBoost settings and complete matched results.

Online Resource 5 (ESM_5.zip): all three exploratory development rounds, source-specific results and source-file amplitude diagnosis (Fig. S4).

References

[1] Radoš K, Brkić M, Begušić D (2024) Recent Advances on Jamming and Spoofing Detection in GNSS. *Sensors* 24(13):4210. https://doi.org/10.3390/s24134210

[2] Li S, Tang X, Lin H, Wang F (2025) GNSS spoofing detection based on frequency domain processing. *Measurement* 242:115872. https://doi.org/10.1016/j.measurement.2024.115872

[3] Mao W, Ni S, Ren J, Zhou W (2026) GNSS spoofing detection based on Q-channel Correlation Signal Quality Monitoring. *Measurement* 258:119458. https://doi.org/10.1016/j.measurement.2025.119458

[4] Foruhandeh M, Mohammed AZ, Kildow G, Berges P, Gerdes R (2020) Spotr: GPS Spoofing Detection via Device Fingerprinting. *WiSec*, pp 242–253. https://doi.org/10.1145/3395351.3399353

[5] Khan N, Marata L, Nurmi J, Lohan ES (2026) GNSS-Spoofing Detection via Machine Learning With In-Domain and Cross-Domain Testing and Open-Access I/Q Data. *IEEE Access* 14:73921–73934. https://doi.org/10.1109/ACCESS.2026.3693081

[6] Tong X, Feng Y, Li JJ (2018) Neyman-Pearson classification algorithms and NP receiver operating characteristics. *Science Advances* 4(2):eaao1659. https://doi.org/10.1126/sciadv.aao1659

[7] Tong X, Xia L, Wang J, Feng Y (2020) Neyman-Pearson classification: parametrics and sample size requirement. *Journal of Machine Learning Research* 21(12):1–48. https://jmlr.org/papers/v21/18-577.html

[8] Foruhandeh M, Mohammed AZ, Kildow G, Gerdes R, Berges P (2021) SatGrid Dataset, realtime genuine and spoofing traces of GPS signals collected at different geographical locations, times and environmental conditions. Virginia Tech, version 2. https://doi.org/10.7294/SE62-7X13

[9] Zhang B, Zhang Q, Wang X, Fu X, Luo H, Qiu T (2026) A Time-Aware Evaluation Framework for Real-Time GPS Spoofing Detection in UAV-Based Power System Monitoring. *IEEE Access*. https://doi.org/10.1109/ACCESS.2026.3704424

[10] Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 785–794. https://doi.org/10.1145/2939672.2939785

[11] Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

[12] Breiman L (2001) Random Forests. *Machine Learning* 45:5–32. https://doi.org/10.1023/A:1010933404324

[13] Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. *Neural Computation* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[14] Gorishniy Y, Kotelnikov A, Babenko A (2025) TabM: Advancing tabular deep learning with parameter-efficient ensembling. *ICLR*. https://proceedings.iclr.cc/paper_files/paper/2025/hash/c1ba41c694834aeef91ae161711d4939-Abstract-Conference.html

[15] Ye H-J, Yin H-H, Zhan D-C, Chao W-L (2025) Revisiting Nearest Neighbor for Tabular Data: A Deep Tabular Baseline Two Decades Later. *ICLR*. https://proceedings.iclr.cc/paper_files/paper/2025/hash/ecd294194d5155432aab533f7ce0f6fe-Abstract-Conference.html

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM5.zip
- ESM1.zip
- ESM2.zip
- ESM3.zip
- ESM4.zip