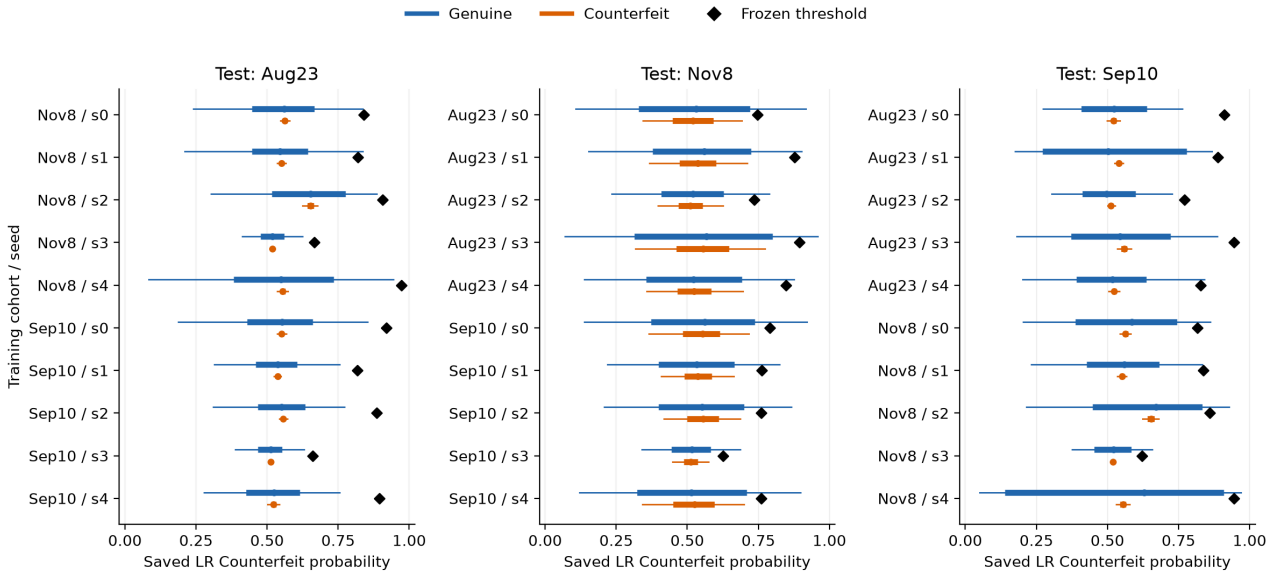
Online Resource 2 Saved score diagnostics

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

Annals of Telecommunications
Bo Xie and Yi Zhao
Wuhan Digital Engineering Institute, Wuhan, China
Corresponding author: Bo Xie (xiebo@mail.hfut.edu.cn)

These analyses were specified after inspecting the main results. They explain archived outputs; they do not select models, change thresholds, or replace the frozen primary analysis. All 270 score blocks and 3,240 calibrated cells were checked. The figures display the same middle allocation used in the main text (k=20, c=160, heldout calibration, nominal 5%). No raw scores are pooled across fitted models.

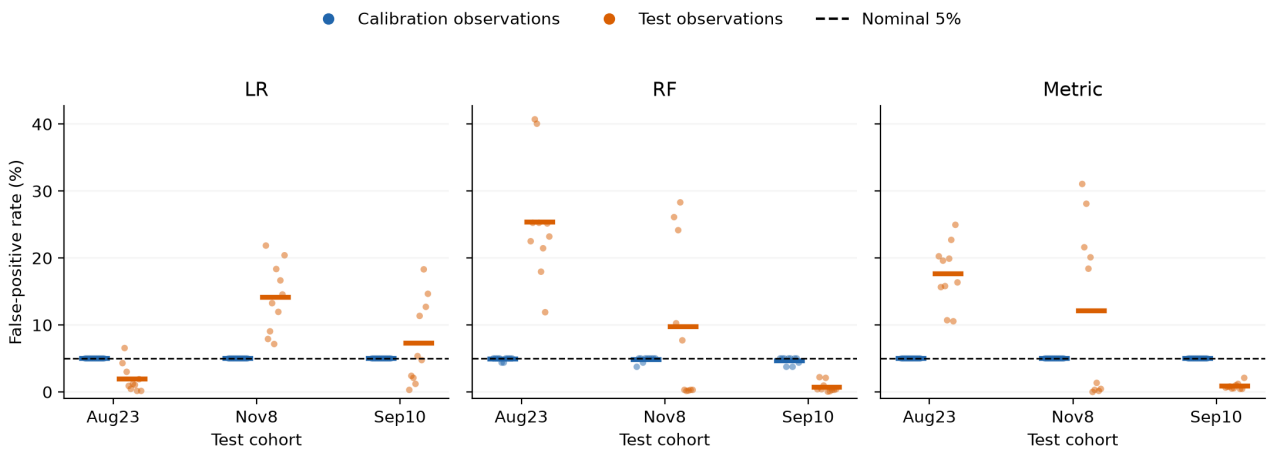
Figure S2. LR score intervals and fixed thresholds



Post hoc: k=20, c=160, heldout calibration, nominal 5%. Thin: 5–95%; thick: 25–75%; dot: median.

Each row is one training-cohort/seed combination evaluated on the panel's test cohort, with calibration supplied by the remaining cohort. All 30 LR blocks for this setting are shown. Thin lines span the within-block 5th-95th percentiles, thick lines span the 25th-75th percentiles, and dots mark medians. Black diamonds are the original thresholds. These are distribution summaries, not confidence intervals. Concentrated Counterfeit scores commonly occupy the interior of the wider Genuine score distribution, below the upper-tail thresholds.

Figure S3. Calibration and test false-positive rates



Post hoc: k=20, c=160, heldout calibration. Each dot: one role/seed block; bars: descriptive means, not CIs.

Each point represents one of the ten role/seed blocks per method and test cohort. Horizontal colored bars are arithmetic means, not confidence intervals; the dashed line marks the nominal 5% level. All 90 method/role/seed blocks for this setting are included. Calibration FPR is evaluated on the observations used to choose the threshold. Agreement with its finite-sample bound therefore does not establish control on the held-out test cohort. RF ties can reduce calibration FPR below 5% under the strict greater-than rule.

Fixed-score quality sensitivity

For this post-hoc sensitivity only, test observations with CN0=0 were omitted from the evaluation denominator. Models, support sets, calibration sets, scores, and thresholds were held fixed. Across the 3,240 cells, the largest absolute changes were 0.173332 percentage points in FPR and 0.136774 points in FNR. This is not a retraining or recalibration analysis and cannot establish that all measurement-quality effects are negligible. The 20 distinct zero-CN0 test observations remain included in every primary result.

Replay and provenance

The 45 LR model states and their support-fitted scalers reproduced the 90 archived LR test score blocks exactly at archived float32 precision. Cell thresholds and error rates were independently recomputed. No fitting calls were made. Support-resubstitution AUROC is diagnostic only and is never presented as generalization performance.