

NA-CoLR: A Noise-Aware, Coherent and Self-Auditing Likelihood-Ratio Engine for Forensic Event Attribution Under Erroneous Reference Labels

Gaseb Alotibi

Department of Computer and Information Technology,
University of Tabuk, Tabuk, Saudi Arabia

Abstract—Forensic examiners are increasingly expected to report the weight of digital evidence as a likelihood ratio (LR) rather than as a class decision. Three obstacles block this transition: calibration methods assume that reference labels are correct, although forensic reference labels are reconstructed and erroneous; multiclass problems are handled by independently calibrated pairwise systems whose reports contradict one another; and the empirical lower and upper bound on reportable LRs is estimated from the same erroneous labels, while reported LRs carry no interval. This paper presents NA-CoLR, a four-stage model that couples a stability-gated estimate of the label-transition matrix, forward-corrected coherent calibration on the probability simplex, a noise-aware and order-preserving bound, and a bootstrap that propagates transition-matrix uncertainty into per-event intervals and a tiered abstention policy. On NSL-KDD, CICIDS2017 and UNSW-NB15 with 15% injected reference-label error, the complete model attains a multiclass log-likelihood-ratio cost of 0.047, 0.036 and 0.354, against 0.170, 0.087 and 0.355 for the same calibrator without correction and 0.092, 0.043 and 0.344 for a one-versus-rest baseline. Correction helps under every noise process tested; the lead over the one-versus-rest baseline holds under pair-flip noise and prevalence shift, but not under instance-dependent noise. Pairwise systems violate transitivity by up to 1.94 decades, whereas NA-CoLR is coherent by construction, including after bounding. Selecting a system by validation against erroneous labels reverses the true ranking and selects a system 4.0 times worse than the best. Unseen classes are not detected reliably, so the model is restricted to a closed, validated class taxonomy.

Keywords—Likelihood ratio; calibration; log-likelihood-ratio cost; label noise; empirical lower and upper bound; evidence evaluation; digital forensics; uncertainty quantification

I. INTRODUCTION

An examiner who reconstructs an event from network flows, audit trails, transaction logs or redo records must eventually state how strongly that reconstruction supports one proposition over another. Driven by ENFSI guidance [1], the National Research Council report [2] and two decades of work in speaker, fingerprint and DNA comparison [3]–[5], the accepted answer is a likelihood ratio (LR) rather than a bare classification. Digital forensics has largely not made that transition. Published models report accuracy, precision and F_1 [6]–[8]; almost none report the log-likelihood-ratio cost C_{llr} , and none report whether the LRs they would produce are calibrated. Attempting the transition exposes three problems.

A. Problem 1: The Reference Labels are Wrong

A forensic corpus is not annotated; it is *reconstructed*, and reconstruction errs. CICIDS2017 contains labelling errors large enough to change reported detection performance [9]; surveys of intrusion datasets identify label quality as a primary limitation [10]; and label-noise mitigation for intrusion detection is an active topic [11], [12]. Calibration theory, however, assumes that validation labels are the truth: C_{llr} is an expectation over trials whose membership is taken as known [13]. When part of that membership is wrong, the resulting C_{llr} is biased in a direction that depends on the calibrator, so the bias does not cancel when two systems are compared. No forensic calibration method that estimates or corrects for this was found, with the caveat on search scope stated in Section II-E.

B. Problem 2: Multiclass LR Systems are Incoherent

Forensic LR theory is predominantly binary. With K candidate event classes, practice fits $K(K-1)/2$ pairwise systems, each calibrated on its own subset of the reference data. The resulting LRs do not derive from a common posterior, so transitivity fails: $LR_{ij} \cdot LR_{jk} \neq LR_{ik}$. Two reports on the same evidence can then form a chain that is internally contradictory. Section VI-F shows that the effect is large: a mean violation of 1.94 decades for the worst system measured, and more than one decade on 81.6% of class triples for the pairwise logistic baseline on NSL-KDD.

C. Problem 3: Bounding is Label-Blind, Uncertainty is Absent

The empirical lower and upper bound (ELUB) [14] exists because sparse tails produce unsupportably extreme LRs. ELUB is estimated from the same erroneous reference data, so its support counts inherit the error; Proposition 2 states exactly when this loosens the bound. Separately, a reported LR is a point value: the examiner cannot say how much it would move had the reconstruction erred differently, and has no principled basis for declining to report, an omission with direct consequences for admissibility [15], [16].

D. Contributions

- NA-CoLR, a model in which noise correction, coherent multiclass calibration, noise-aware bounding and uncertainty quantification are coupled: Stages 2–4 all

consume the transition matrix estimated and gated by Stage 1.

- A coherent calibration layer for which all pairwise LR's are transitive by construction (Proposition 1).
- A common-floor NA-ELUB (Section IV-E) that bounds unsupported evidence while preserving transitivity and class order and leaving supported LR's unchanged (Proposition 3), with an explicit condition under which noise loosens the naive bound (Proposition 2).
- A stability-based shrinkage criterion that decides how far to trust an estimated transition matrix without a clean label.
- A bootstrap that re-estimates the transition matrix in every replicate, yielding per-event LR intervals, a C_{lr}^K band and a tiered report/conservative/abstain policy, together with an operational protocol for diagnostics, provenance and human review (Section VII).
- A demonstration that validating LR systems against erroneous references inverts the ranking of candidates under C_{lr} (Section VI-I), with the mechanism that makes the reversal systematic.

II. RELATED WORK

A. Likelihood Ratios and Their Validation

The LR framework for evidence evaluation is set out in [3], [17], [18], with forensic validation guidance in [1], [5], [19]. The cost C_{lr} originates in speaker recognition [13] and decomposes into discrimination and calibration loss through the pool-adjacent-violators algorithm [13], [20]; the underlying decomposition of probabilistic accuracy is older still [21]. Tippet, ECE and DET plots are standard companions [20], [22]. Later work refines the metric family with devPAV [23], proposes bi-Gaussianised calibration [24], questions whether C_{lr} alone suffices as a calibration test [25], and extends score-based LR's to new modalities [26], [27]. Reference datasets for validation exist [28]. Throughout this literature the reference labels are treated as exact.

B. Learning Under Label Noise

Consistent estimation under class-conditional noise is given in [29]; forward and backward loss correction with a transition matrix T in [30]; anchor-free and bilevel estimators of T in [31]–[33]. Sample-selection and co-training approaches include [34], [35], and robust losses are surveyed in [36], [37]. Security-specific treatments exist for software defects [38] and intrusion detection [11], [12]. This work targets classification accuracy; none of it produces a reportable weight of evidence or examines what label noise does to a *calibrated* LR.

C. Calibration and Uncertainty

Confidence calibration [39]–[41] and its multiclass extension [42] supply the machinery adapted here, and distribution-free uncertainty is available through conformal prediction [43], [44]. Calibrated posteriors are not LR's, however: a posterior embeds a prior that a forensic examiner must not assume. The

intervals in this paper are obtained by bootstrap resampling [45] of the whole calibration side, including the transition-matrix estimate.

D. Digital and Database Forensics

Metamodels and process models for database forensics are developed in [46]–[49], with an account of the wider subdomain landscape in [50] and an applied treatment in [51]; dataset scarcity is documented in [52], [53]. Learning-based detection over audit and flow data is surveyed in [6]–[8], [54]–[56], with recent architectures in [57]–[60]. Explainability requirements related to admissibility are treated in [61]–[66], typically via feature attribution [67], and the consequences of unreliable algorithmic evidence in legal settings in [16], [68]. None of this work reports LR's or treats reference error as a calibration problem.

E. Positioning and Limits of this Survey

The gap addressed here is an intersection: label-noise robustness has been developed for classifiers rather than for LR reporting [36], [69], [70]; forensic LR calibration assumes exact references; and multiclass coherence of forensic LR reports appears not to have been examined. The literature search was targeted, not systematic (no PRISMA protocol, database enumeration or screening), and statements of absence should be read at that strength. Forward correction [30], simplex calibration [42] and bootstrap intervals [45] are established, and Proposition 1 is elementary. What is new is the common-floor bound, the condition of Proposition 2, the coupling of the four stages, and the measurement of what reference error costs a forensic LR system.

III. PROBLEM FORMALISATION

Let $x \in \mathcal{X}$ be an event's feature vector, $y \in \{1, \dots, K\}$ its true class, and $\tilde{y}$ the observed reference label. Class-conditional noise is characterised by $T \in [0, 1]^{K \times K}$ with $T_{ij} = P(\tilde{y} = j \mid y = i)$. For propositions H_i ($y = i$) and H_j ($y = j$),

$$LR_{ij}(x) = \frac{p(x \mid y = i)}{p(x \mid y = j)}. \quad (1)$$

The class taxonomy $\{1, \dots, K\}$ is assumed closed: every event belongs to one of the K validated classes. Section VI-L tests what happens when it does not.

A. Multiclass Log-LR Cost

C_{lr} is extended to K classes as the prior-free mean over unordered class pairs:

$$C_{lr}^K = \frac{2}{K(K-1)} \sum_{i < j} \frac{1}{2} \left[\frac{1}{n_i} \sum_{m: y_m = i} \log_2 \left(1 + \frac{1}{LR_{ij}(x_m)} \right) + \frac{1}{n_j} \sum_{m: y_m = j} \log_2 \left(1 + LR_{ij}(x_m) \right) \right]. \quad (2)$$

Eq. (2) reduces to the standard C_{lr} when $K = 2$: 0 is perfect, 1 is uninformative, and values above 1 indicate

TABLE I. STAGE INTERFACES OF NA-CoLR. n : NUMBER OF EVENTS; K : NUMBER OF CLASSES. EVERY CHECK THAT FAILS RAISES AN ERROR AND NO LR IS REPORTED

Stage	Input	Output	Checks and handling of invalid input
0. Base model	features X ; noisy training labels $\tilde{y}_{tr}$	log-posteriors $z \in \mathbb{R}^{n \times K}$	rows with missing features excluded upstream; z must have K columns, no NaN; un-normalised logits renormalised by log-sum-exp; probabilities floored at 10^{-9} before the logarithm and the count of floored events recorded.
1. Transition	$z_{val}, \tilde{y}_{val}; q, d_0$	$\hat{T}$ (row-stochastic), $\lambda \in [0, 1], \hat{T}_\lambda$	labels must be integers in $[0, K)$; every class needs ≥ 20 validation events (otherwise merge or remove the class); a predicted class with < 20 events receives an identity row; condition number of $\hat{T}_\lambda$ recorded.
2. Calibration	$z_{val}, \tilde{y}_{val}, \hat{T}_\lambda$; training class counts	parameters $(a, b) \in \mathbb{R}^{2K}$; scores $s(x) \in \mathbb{R}^K$ by (8)	log-softmax evaluated stably; forward probabilities floored at 10^{-12} ; class proportions π_i must be positive.
3. Bound	$s, \tilde{y}_{val}, \hat{T}_\lambda$	bounded scores s' ; $\log_{10} \text{LR}_{ij} = s'_i - s'_j$	deconvolved counts clipped at 1; if $\hat{T}_\lambda$ is singular the observed counts are used; s' must be finite.
4. Uncertainty	$z_{val}, \tilde{y}_{val}, z_{case}; B, \alpha, \tau$	interval $[\ell_{ij}, h_{ij}]$ and tier per event and pair	replicate λ range recorded; tier rules of Section IV-F.

a system worse than reporting nothing. Because each class is weighted by $1/n_i$, C_{lr}^K does not depend on the class proportions of the test set.

B. Coherence

Definition 1 (Coherence). A set of LRs $\{\text{LR}_{ij}(x)\}$ is coherent if for all i, j, k and all x , $\text{LR}_{ij}(x) \cdot \text{LR}_{jk}(x) = \text{LR}_{ik}(x)$ and $\text{LR}_{ij} = 1/\text{LR}_{ji}$.

Incoherence is measured as the mean absolute violation in decades, $\mathbb{E} |\log_{10}(\text{LR}_{ij}\text{LR}_{jk}/\text{LR}_{ik})|$, together with the fraction of triples on which it exceeds one decade, since a one-decade shift moves a report across a verbal-scale boundary [1].

C. Distance From the True LR

C_{lr} measures whether an LR points the right way with the right confidence *relative to labels*; it does not measure whether the number is correct. Where the data-generating process is known, the true LR_{ij}^* is available and the *Analytic Log-LR Deviation* is

$$\text{ALRD} = \sqrt{\frac{2}{K(K-1)n} \sum_m \sum_{i < j} \left(\log_{10} \frac{\text{LR}_{ij}(x_m)}{\text{LR}_{ij}^*(x_m)} \right)^2}, \quad (3)$$

restricted to pairs with $|\log_{10} \text{LR}^*| \leq 6$, the top of the ENFSI verbal scale. The *overstatement rate* is the fraction of these pairs on which $|\log_{10} \text{LR}|$ exceeds $|\log_{10} \text{LR}^*|$ with the same sign.

IV. THE NA-CoLR MODEL

A. Architecture and Stage Interfaces

NA-CoLR is a four-stage pipeline. Stage 1 estimates the transition matrix and decides how far to trust it; Stages 2–4 consume that decision. Table I specifies what each stage receives and emits and which checks stop processing; the released module `pipeline.py` implements these contracts.

B. Stage 1A: Estimating the Transition Matrix

A gradient-boosted tree ensemble is fitted to the *noisy* training labels, producing $\tilde{p}(\cdot | x)$. When the base model is far more accurate than the noise rate, its confident predictions act as pseudo-clean labels, which motivates the confidence-filtered confusion estimator:

$$\hat{T}_{ij} \propto |\{m \in \mathcal{A}_i : \tilde{y}_m = j\}|, \quad (4)$$

where, $\mathcal{A}_i$ is the top- q fraction of validation events by confidence among those predicted as class i . With $q = 1$ (no filter) the mean row L_1 error of $\hat{T}$ on UNSW-NB15 is 0.71; with $q = 0.20$ it is 0.38, and 0.02 on the two datasets where the base model is strong. The estimator assumes class-conditional noise; Section VI-D tests it under noise that is not.

C. Stage 1B: How Far to Trust $\hat{T}$

Forward correction with a badly estimated $\hat{T}$ is harmful: on UNSW-NB15 it raises C_{lr}^K from 0.355 to 0.493. $\hat{T}$ is therefore shrunk toward the identity, $\hat{T}_\lambda = (1 - \lambda)I + \lambda\hat{T}$, with λ selected by the *stability* of the estimator. T is re-estimated on two disjoint halves of the validation split, giving $\hat{T}_a, \hat{T}_b$, and

$$d = \frac{1}{K} \|\hat{T}_a - \hat{T}_b\|_1, \quad \lambda = \text{clip}\left(1 - \frac{d}{d_0}, 0, 1\right), \quad (5)$$

with $d_0 = 0.45$. Selecting λ by held-out corrected likelihood is tautological (fitting and scoring with $\hat{T}$ always returns $\lambda = 1$), and selecting it by C_{lr} against the reference labels is anti-correlated with true performance (Section VI-I). Criterion (5) uses only the agreement of the estimator with itself. Its sensitivity to q and d_0 is reported in Section VI-D.

D. Stage 2: Coherent Calibration on the Simplex

Let $z(x) = \log \tilde{p}(\cdot | x) \in \mathbb{R}^K$. The calibration map is a single affine transform on the simplex,

$$\hat{p}(y = i | x) \propto \exp(a_i z_i(x) + b_i), \quad (6)$$

with $(a, b) \in \mathbb{R}^{2K}$ fitted on validation by minimising the forward-corrected negative log likelihood

$$\mathcal{L}(a, b) = -\frac{1}{n} \sum_m \log \left([\hat{T}_\lambda^\top \hat{p}(\cdot | x_m)]_{\hat{y}_m} \right) \quad (7)$$

plus an ℓ_2 penalty of 10^{-6} (L-BFGS-B, at most 500 iterations). Under a correctly specified $\hat{T}_\lambda$ the minimiser of (7) recovers the clean posterior [30]. Removing the prior via the training class proportions π gives prior-free per-class scores in decades,

$$s_i(x) = \frac{\log \hat{p}(y = i | x) - \log \pi_i}{\log 10}, \quad \log_{10} \text{LR}_{ij} = s_i - s_j. \quad (8)$$

Proposition 1 (Coherence). *For LR_s defined by (8) from any single score vector $s(x)$, the set $\{\text{LR}_{ij}(x)\}$ is coherent.*

Proof: $\log_{10} \text{LR}_{ij} + \log_{10} \text{LR}_{jk} = (s_i - s_j) + (s_j - s_k) = s_i - s_k = \log_{10} \text{LR}_{ik}$, and antisymmetry is immediate. ■

The pairwise construction admits no such cancellation, because its $K(K-1)/2$ calibrators have independent parameters fitted on different data subsets. Any construction that differences one score vector, including one-versus-rest calibration, is coherent in the same way.

Numerical specification. Base-model probabilities are floored at 10^{-9} before the logarithm, so z is finite; the floor was reached by 0%, 0% and 0.36% of test events on the three datasets. The log-softmax in (6) is evaluated with the log-sum-exp identity, and $\hat{T}_\lambda^\top \hat{p}$ is floored at 10^{-12} inside (7). Unbounded scores can still be extreme: the largest $|\log_{10} \text{LR}|$ after Stage 2 was 19.0, 26.5 and 10.3 decades. Stage 3 limits every reported value to at most $U_m \leq 2\log_{10}(n_{\max}+1)$, which was 7.3, 6.9 and 5.6 decades. No non-finite value occurred in any experiment, and `pipeline.py` raises an error if one does.

E. Stage 3: Noise-Aware ELUB

Let $u_i = \log_{10}(n_i + 1)$ be the empirical support of class i , where the counts are first *deconvolved*, $n \leftarrow \hat{T}_\lambda^{-\top} \tilde{n}$, clipped at 1. For each event, with $U_m = u_{(1)} + u_{(2)}$ the supports of its two highest-scoring classes, the bound sets a common floor:

$$s'_i = \max(s_i, \max_k s_k - U_m). \quad (9)$$

The deconvolution step is what makes the bound noise-aware; Proposition 2 gives the exact condition under which it matters.

Proposition 2 (When noise loosens the naive bound). *Let n be the true class counts of the reference set. Then $\mathbb{E}[\tilde{n}] = T^\top n$, and the naive support of class i , evaluated at its expected count, exceeds the true support if and only if*

$$\sum_{k \neq i} n_k T_{ki} > n_i(1 - T_{ii}), \quad (10)$$

i.e. the inflow of mislabelled events exceeds the outflow. If $T^\top n \neq n$, at least one class satisfies (10). If $T^\top n = n$, which

holds for example for any doubly stochastic T on balanced classes, no class is inflated in expectation.

Proof: $\mathbb{E}[\tilde{n}_i] = \sum_k n_k T_{ki}$ by definition, and $\mathbb{E}[\tilde{n}_i] > n_i$ rearranges to (10). Because each row of T sums to one, $\sum_i \mathbb{E}[\tilde{n}_i] = \sum_i n_i$; two vectors with equal sums that differ have a component in which the first is larger. Monotonicity of $\log_{10}(\cdot + 1)$ completes the proof. ■

The effect is therefore class-specific, not systematic: noise inflates the support of classes that gain mass and deflates that of classes that lose it. Solving $\hat{T}_\lambda^\top n = \tilde{n}$ is unbiased when $\hat{T}_\lambda = T$, so its goal is an *accurate* support, not a uniformly tighter one. On NSL-KDD, (10) predicts inflation of Probe and R2L; their observed counts were 1,678 and 387 against true counts of 996 and 275, and deconvolution gave 1,013 and 294. Over all classes, the largest support error fell from 0.23 to 0.03 decades on NSL-KDD and from 0.41 to 0.19 decades on CICIDS2017. On UNSW-NB15, $\lambda = 0$ disables deconvolution, and the Shellcode support remains inflated by 0.37 decades.

Proposition 3 (Common-floor bound). *For s' defined by (9): 1) $\{\text{LR}_{ij}\}$ computed from s' is coherent; 2) $s_i \geq s_j \Rightarrow s'_i \geq s'_j$, so no reported LR changes sign; 3) $\max_{ij} |s'_i - s'_j| \leq U_m$; and 4) every LR between the leading class and a class with $\max_k s_k - s_j \leq U_m$ is unchanged.*

Proof: 1) follows from Proposition 1. 2) The same non-decreasing map $t \mapsto \max(t, c_m)$ is applied to every coordinate. 3) The largest coordinate is unchanged and every coordinate is at least $\max_k s_k - U_m$. 4) Neither coordinate is altered. ■

Pairwise clipping in LR space has no single score vector and loses transitivity (1.41 decades on NSL-KDD, Table VI); per-class clipping to $\pm u_i$ can reorder classes, and on a deliberately imbalanced five-class construction changes the sign of 17,654 of 200,000 reported LR_s. A scalar shrinkage $s' = c + \lambda_m(s - c)$ also preserves order and coherence, but one extreme gap between two implausible classes then compresses the informative gap between the two leading classes; the ablation (Table VI, “v1 shrinkage”) shows that this raises C_{lr}^K from 0.036 to 0.111 on CICIDS2017. The floor touches only the scores that exceed the support.

F. Stage 4: Bootstrap Intervals and Tiered Abstention

For $b = 1, \dots, B$, validation events are resampled with replacement, and Stages 1a, 1b, 2 and 3 are all re-run on the resample, so every replicate re-estimates $\hat{T}$ and λ . The $\alpha/2$ and $1 - \alpha/2$ quantiles of the replicate LR_s give $[\ell_{ij}, h_{ij}]$; the same replicates evaluated with C_{lr}^K give a band. Only the calibration side is refitted, so the cost is B small convex optimisations.

Each reported pair is assigned a tier. ABSTAIN: the interval contains LR = 1, so the direction of support is undetermined. CONSERVATIVE: the interval excludes 1 but is wider than $\tau = 2$ decades; the end nearer to 1 is reported as a lower bound (“the evidence is at least 10^ℓ times more probable under H_i ”). REPORT: the interval excludes 1 and is at most τ wide; the point value is reported with its interval. The pair that decides the attribution of an event is the leading class against the runner-up; if that pair abstains, the event is reported as *inconclusive* and is referred for examination by other means (Section VII).

TABLE II. DATASETS AFTER PREPROCESSING

Dataset	K	Events	Features	Test
NSL-KDD [71]	4	70,000	41	10,500
CICIDS2017 [72]	6	56,625	77	8,494
UNSW-NB15 [73], [74]	9	70,000	42	10,500

G. The Complete Procedure

Algorithm 1 states NA-CoLR end to end. The dominant cost is one base-model fit plus $B + 2$ convex optimisations in $2K$ parameters; for $K \leq 9$ and $n \leq 70,000$ the complete procedure with $B = 30$ runs in under four minutes on one CPU core.

Algorithm 1 NA-CoLR

Input: splits with noisy labels $\tilde{y}$; q, d_0, B, α, τ

Output: coherent bounded LR, intervals, tiers, provenance record

Validate inputs (Table I) **Stage 1a.** Fit base model on $(X_{tr}, \tilde{y}_{tr})$; $z \leftarrow \log \tilde{p}(\cdot | x)$; estimate $\hat{T}$ by (4) **Stage 1b.** Set λ by (5); $\hat{T}_\lambda \leftarrow (1 - \lambda)I + \lambda\hat{T}$ **Stage 2.** Fit (a, b) by (7); scores s by (8) **Stage 3.** Deconvolve counts with $\hat{T}_\lambda$; bound s by (9) **Stage 4.** **for** $b \leftarrow 1$ **to** B **do**

 Resample validation events; re-run Stages 1a–3; record $\log_{10} \text{LR}^{(b)}$

end

Quantiles give $[\ell, h]$; assign tiers (Section IV-F) Write the provenance record (Table VIII)

V. EXPERIMENTAL SETUP

A. Datasets

Three public benchmarks are used (Table II), obtained from public mirrors whose MD5 checksums are recorded in the released harness. Classes with fewer than 200 instances are removed and each corpus is capped at 70,000 events by stratified subsampling. Categorical features are ordinally encoded and all features standardised. Splits are stratified 70/15/15 with a fixed seed. These are network-flow corpora; they stand in for forensic event attribution in general and are not database artefacts (Section VIII).

B. Reference-Label Error

The main study injects asymmetric class-conditional noise that sends the error mass ε of class i to class $i + 1$, with $\varepsilon = 0.15$, and Section VI-C sweeps $\varepsilon \in \{0, 0.05, 0.10, 0.15, 0.20, 0.30\}$. Because this structure matches the assumption of estimator (4), three further noise processes at the same rate are used in Section VI-D: symmetric noise; random pair-flip noise, in which each class sends its error mass to a randomly chosen other class; and *instance-dependent* noise, in which the flip probability of an event grows with its closeness to the nearest other class centroid and the flip goes to that class. The last violates the class-conditional assumption. Models are trained and calibrated on noisy labels and evaluated against the retained clean labels.

C. The Analytic Oracle Testbed

ALRD requires a known data-generating process. For each dataset, class-conditional diagonal Gaussians are fitted to its twelve most discriminative features (ANOVA F) and a fresh corpus is sampled from the fitted model, so LR^* is exact for every event and pair. The entire pipeline, including noise injection and $\hat{T}$ estimation, is re-run on it.

D. Systems Compared

- B1 pairwise logistic calibration, the de facto multiclass forensic practice; **B2** pairwise PAV [13];
- B3 one-versus-rest logistic; **B4** coherent simplex calibration without noise correction (the matched ablation of Stage 1);
- B5 B1 with naive ELUB [14]; **B6** noise-corrected representation with pairwise calibration;
- NA-CoLR-U Stages 1–2 only; **NA-CoLR** the full model. To attribute gains to components, Section VI-B also applies the correction and the bound to B3.

E. Configuration

Base model: histogram gradient boosting in the family of [75], [76], 180 iterations, learning rate 0.12, early stopping. $q = 0.20$, $d_0 = 0.45$, $B = 30$, $\alpha = 0.05$, $\tau = 2$ decades, seed 20260115. Software: Python with NumPy 2.4.4, SciPy 1.17.1, pandas 3.0.2 and scikit-learn 1.8.0, pinned in the release because results change across library versions (Section VIII).

VI. RESULTS

A. Overall Performance

Table III and Fig. 1 give the comparison at $\varepsilon = 0.15$. Against its matched ablation B4, which differs only in the absence of Stage 1, NA-CoLR lowers C_{lr}^K from 0.170 to 0.047 on NSL-KDD and from 0.087 to 0.036 on CICIDS2017. Against the strongest baseline, B3, it is lower on these two datasets (0.092 and 0.043) and higher on UNSW-NB15 (0.354 against 0.344). On UNSW-NB15 the stability criterion returned $\lambda = 0$ and NA-CoLR reduces to B4 with a bound; without the criterion, correction raises C_{lr}^K to 0.493. The determining factor is base-model accuracy: 99.2% and 99.3% where correction succeeds and 81.5% where it does not. On ALRD, NA-CoLR is the best system on NSL-KDD (2.71 decades against 2.95 for B4 and B3) and within 0.10 decades of the best on the other two datasets. NA-CoLR does not lead on overstatement: its rate (0.26, 0.20, 0.18) exceeds that of B4 (0.18, 0.15, 0.17), because correction sharpens the LR.

B. Which Component is Responsible

B3 is coherent by construction, so the comparison with B3 isolates the calibration head rather than coherence. Table IV applies NA-CoLR's Stage 1 (correction) and Stage 3 (bound) to B3. The bound changes B3's C_{lr}^K by less than 0.001 while capping its largest LR. The correction helps B3 on NSL-KDD (0.092 $\rightarrow$ 0.080) but not on CICIDS2017 (0.043 $\rightarrow$ 0.046) or UNSW-NB15. Two conclusions follow. First, the one-versus-rest head is itself fairly robust to class-conditional noise, so

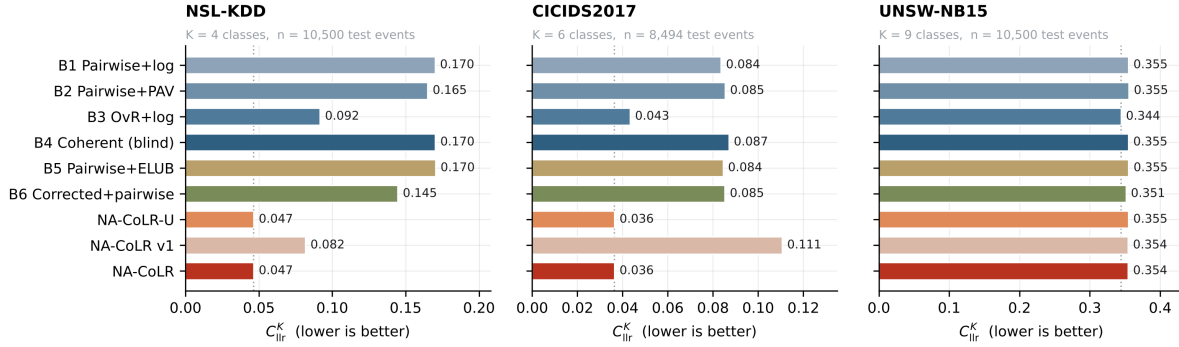


Fig. 1. Multiclass log-likelihood-ratio cost at 15% reference-label error. The dotted line marks the best system on each dataset. NA-CoLR v1 is the scalar-shrinkage bound retained for comparison.

TABLE III. TEST-SPLIT PERFORMANCE AT $\varepsilon = 15\%$ REFERENCE-LABEL ERROR. C_{lr}^K AND ALRD: LOWER IS BETTER; *coh.*: MEAN COHERENCE VIOLATION IN DECADES; *ov.*: OVERSTATEMENT RATE. BEST IN EACH COLUMN IN BOLD.

System	NSL-KDD (K=4)				CICIDS2017 (K=6)				UNSW-NB15 (K=9)			
	C_{lr}^K	ALRD	coh.	ov.	C_{lr}^K	ALRD	coh.	ov.	C_{lr}^K	ALRD	coh.	ov.
B1 pairwise logistic	0.170	2.89	1.57	0.24	0.084	2.45	0.95	0.15	0.355	2.22	0.43	0.16
B2 pairwise PAV	0.165	3.04	0.91	0.32	0.085	2.55	1.23	0.28	0.355	2.31	0.67	0.28
B3 one-vs-rest logistic	0.092	2.95	0.00	0.26	0.043	2.37	0.00	0.22	0.344	2.31	0.00	0.22
B4 coherent, noise-blind	0.170	2.95	0.00	0.18	0.087	2.46	0.00	0.15	0.355	2.32	0.00	0.17
B5 pairwise + naive ELUB	0.170	2.87	0.85	0.23	0.084	2.47	0.57	0.14	0.355	2.22	0.40	0.14
B6 corrected, pairwise	0.145	2.86	1.67	0.20	0.085	2.43	1.94	0.16	0.351	2.27	0.33	0.16
NA-CoLR-U (Stages 1–2)	0.047	5.68	0.00	0.48	0.036	2.42	0.00	0.20	0.355	2.32	0.00	0.18
NA-CoLR v1 (shrinkage bound)	0.082	3.07	0.00	0.28	0.111	2.42	0.00	0.20	0.354	2.32	0.00	0.18
NA-CoLR (full)	0.047	2.71	0.00	0.26	0.036	2.42	0.00	0.20	0.354	2.32	0.00	0.18

TABLE IV. COMPONENT ATTRIBUTION AT $\varepsilon = 15\%$ (C_{lr}^K ; MAXIMUM $|\log_{10} LR|$ IN PARENTHESES)

System	NSL-KDD	CICIDS2017	UNSW-NB15
B4 coherent, noise-blind	0.170 (4.5)	0.087 (5.4)	0.355 (10.3)
B3 one-vs-rest	0.092 (7.0)	0.043 (7.6)	0.344 (7.6)
B3 + bound	0.092 (6.5)	0.043 (6.6)	0.344 (5.8)
B3 + correction	0.080 (7.4)	0.046 (7.6)	0.346 (7.9)
B3 + correction + bound	0.080 (6.4)	0.046 (6.6)	0.345 (6.2)
NA-CoLR-U	0.047 (19.0)	0.036 (26.5)	0.355 (10.3)
NA-CoLR	0.047 (7.3)	0.036 (6.9)	0.354 (5.6)

most of the benefit of correction is realised with the simplex head, which is the head that requires it. Second, the bound and the uncertainty stages are head-agnostic and can be added to B3 at negligible cost; the complete NA-CoLR remains the lowest- C_{lr}^K configuration on two of three datasets.

C. Sensitivity to the Reference-Label Error Rate

Fig. 2 traces C_{lr}^K as ε grows. The slope of a least-squares line through the six points measures how fast the reported weight of evidence degrades with the reconstruction. For NA-CoLR it is 0.335, 0.321 and 0.225 on the three datasets, against 0.874, 0.477 and 0.251 for B4 and 0.671, 0.394 and 0.179 for B3. NA-CoLR is the flattest on NSL-KDD and CICIDS2017 and B3 on UNSW-NB15.

D. Robustness to the Noise Model and to the Settings

Table V repeats the comparison on NSL-KDD and CICIDS2017 under the three alternative noise processes. Cor-

rection lowers C_{lr}^K relative to B4 under every process, by 25–77%. Relative to B3, NA-CoLR is better under random pair-flip noise (0.030 against 0.045; 0.014 against 0.026) and under symmetric noise on NSL-KDD, but worse under symmetric noise on CICIDS2017 and under instance-dependent noise on both datasets (0.092 against 0.077; 0.055 against 0.035). The instance-dependent result is the expected failure of a class-conditional correction: the stability criterion still returned λ of 0.82 and 0.74, because $\hat{T}$ is stable even when it is misspecified. Stability is therefore a necessary condition for trusting $\hat{T}$, not a sufficient one (Section VII).

Over a grid of $q \in \{0.1, 0.2, 0.3, 0.5, 1.0\}$ and $d_0 \in \{0.3, 0.45, 0.6, 0.9\}$, λ ranged over $[0.80, 0.93]$, $[0.62, 0.87]$ and $[0, 0.48]$, and C_{lr}^K over $[0.041, 0.053]$, $[0.023, 0.042]$ and $[0.340, 0.354]$ on the three datasets. Every one of the 20 settings improved on B4 on NSL-KDD and CICIDS2017, so the result does not depend on the chosen q and d_0 .

E. Class Imbalance and Prevalence Shift

Two kinds of shift were tested. In the first, the development (training and validation) data were subsampled to Dirichlet-drawn class proportions, giving total-variation distances of 0.22–0.50 from the test proportions, while the test set was unchanged. NA-CoLR had the lowest C_{lr}^K of the three systems in all six draws (Table V), although its cost rose with the shift on NSL-KDD (0.067 to 0.109). In the second, the fitted systems were evaluated on ten test sets resampled to Dirichlet-drawn proportions. Because (2) weights each class equally, C_{lr}^K moved little: mean 0.047 (s.d. 0.002), 0.035 (0.008) and 0.353 (0.013). Prevalence enters the reported LR only through π in

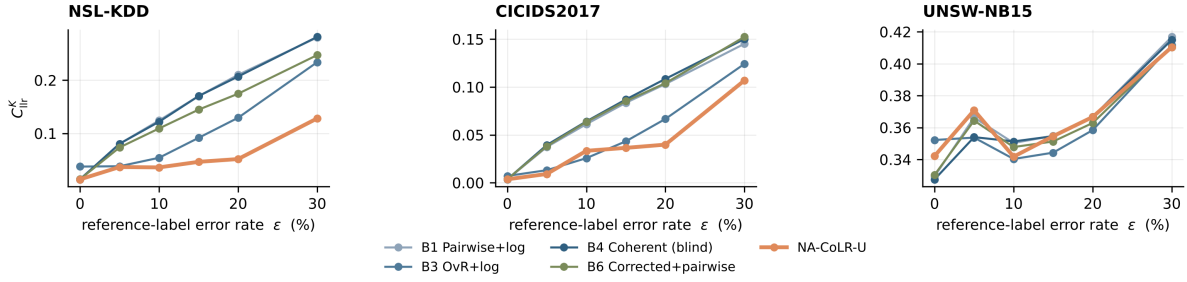


Fig. 2. C_{illr}^K as reference labels degrade. The UNSW-NB15 panel is non-monotone; its classes are intrinsically confusable and the base model reaches only 81.5% accuracy.

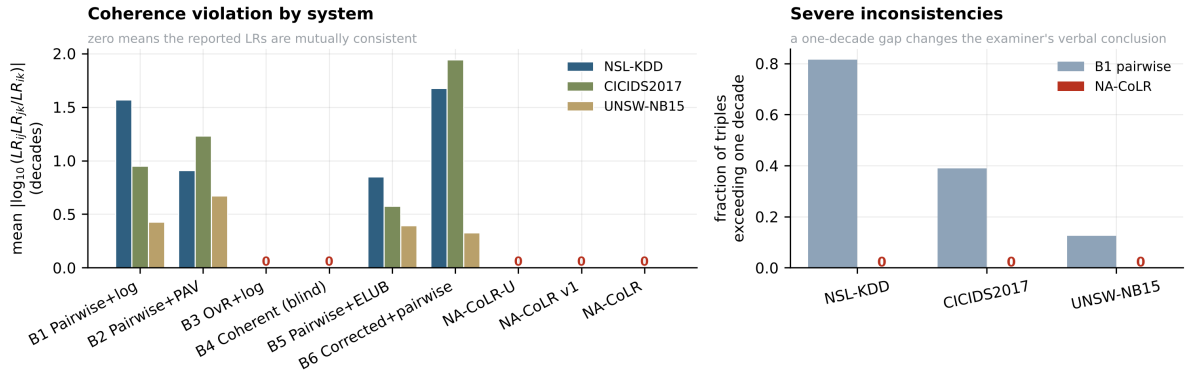


Fig. 3. Coherence violations. Systems derived from a single score vector (B3, B4, NA-CoLR) are coherent by construction.

TABLE V. C_{illr}^K UNDER ALTERNATIVE NOISE PROCESSES ($\varepsilon = 15\%$) AND UNDER DEVELOPMENT PREVALENCE SHIFT (TV : TOTAL-VARIATION DISTANCE BETWEEN DEVELOPMENT AND TEST CLASS PROPORTIONS)

Condition	NSL-KDD			CICIDS2017		
	B4	B3	NA-CoLR	B4	B3	NA-CoLR
asymmetric ($i \rightarrow i+1$)	0.170	0.092	0.047	0.087	0.043	0.036
symmetric	0.179	0.071	0.068	0.102	0.033	0.043
random pair-flip	0.103	0.045	0.030	0.059	0.026	0.014
instance-dependent	0.144	0.077	0.092	0.073	0.035	0.055
shift, $TV=0.24 / 0.22$	0.148	0.090	0.068	0.092	0.054	0.052
shift, $TV=0.36 / 0.34$	0.143	0.086	0.067	0.087	0.053	0.036
shift, $TV=0.50 / 0.40$	0.163	0.121	0.109	0.079	0.044	0.041

(8); if deployment prevalence differs, the LR remains valid, but posterior statements must use case-specific priors supplied by the court, not π .

F. Coherence

Fig. 3 quantifies Problem 2. Every pairwise system contradicts itself: mean violations of 1.57 (B1) and 1.67 (B6) decades on NSL-KDD and 1.94 (B6) on CICIDS2017, and on B1/NSL-KDD 81.6% of class triples are inconsistent by more than one decade. NA-CoLR, B3 and B4 are coherent to machine precision, and NA-CoLR remains coherent after bounding, whereas naive pairwise clipping reintroduces violations of 1.41 and 1.81 decades. Coherence is a property of the construction, not of NA-CoLR specifically; the finding is that the construction standard practice uses lacks it.

G. Bounding and the Distance from the Truth

Fig. 4 plots ALRD against C_{illr}^K . On NSL-KDD, the bound lowers ALRD from 5.68 to 2.71 decades at no cost in C_{illr}^K (0.0465 in both cases), whereas the scalar shrinkage of v1 cost 0.035. On CICIDS2017 and UNSW-NB15, all systems lie within 2.22–2.55 decades and ALRD does not separate them: where the discriminative model is near the generative optimum, the residual is dominated by the mismatch between a boosted-tree posterior and a Gaussian likelihood. ALRD is therefore treated as a diagnostic, not a ranking metric.

H. Tippett Plots

Fig. 5 gives Tippett plots for the best-supported class pair. On NSL-KDD, NA-CoLR separates the two populations by more than the pairwise baseline, and the vertical segment near 6.8 decades shows the bound acting on events whose LR exceeds the support. On CICIDS2017 the upper curves coincide and NA-CoLR moves the lower curve further from zero. On UNSW-NB15 the systems nearly coincide, as expected given $\lambda = 0$.

I. The Model-Selection Inversion

Fig. 6 reports the result with the widest practical reach. The underlying phenomenon is known for classification: correcting label errors in benchmark test sets reorders model rankings [77], and label noise degrades calibration [36], [69]. The present result concerns LR reporting under C_{illr} , the metric forensic validation uses, and holds independently of NA-CoLR. Sweeping the correction strength λ on NSL-KDD,

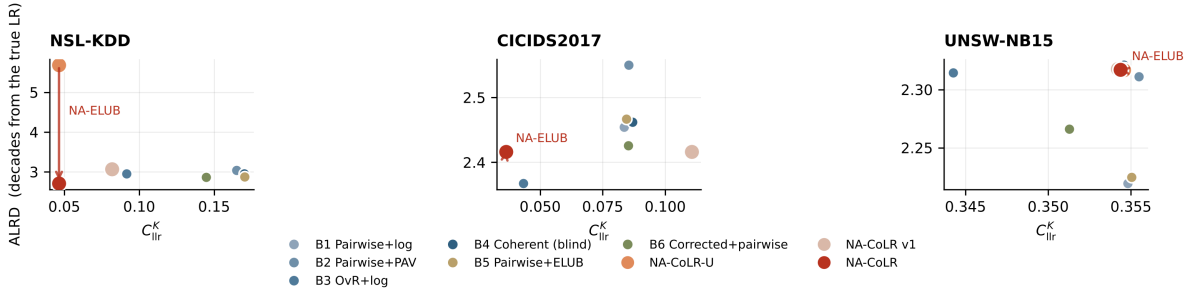


Fig. 4. ALRD against C_{lr}^K . The arrow marks the effect of the bound on NA-CoLR-U.

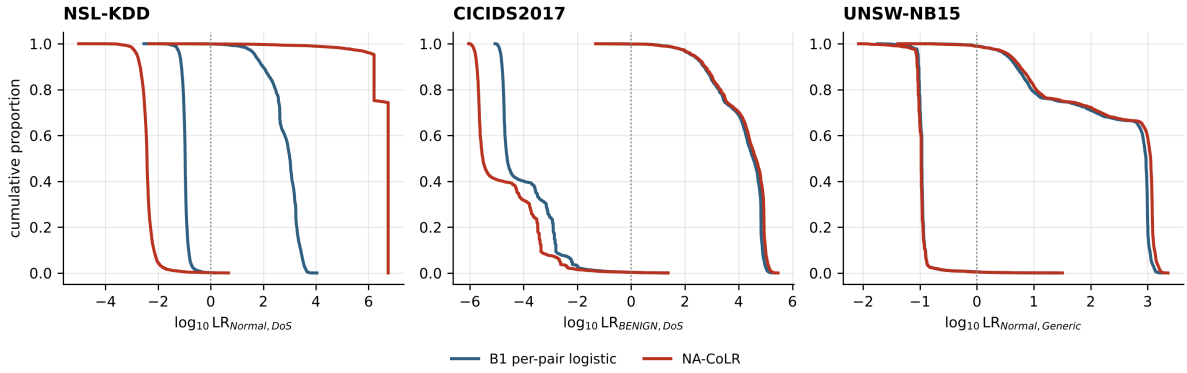


Fig. 5. Tippet plots for the best-supported class pair at $\varepsilon = 15\%$. Curves to the right of zero: events of the first class; to the left: events of the second class. Less mass on the wrong side of zero is better.

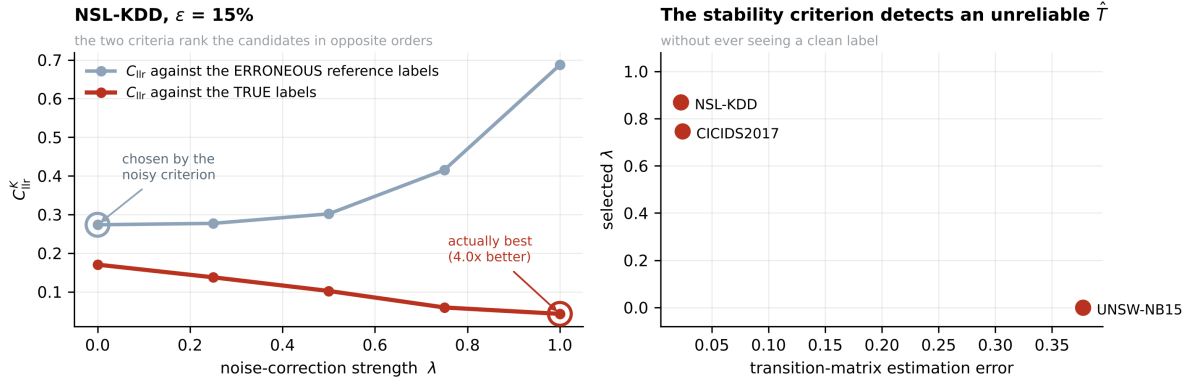


Fig. 6. Left: Validating against erroneous references selects the worst available system. Right: The selected λ against the actual transition-matrix error; The criterion disengages where $\hat{T}$ is unreliable without using clean labels.

C_{lr}^K against the *erroneous* test labels rises monotonically from 0.273 to 0.687, while C_{lr}^K against the *true* labels falls monotonically from 0.170 to 0.043. The two criteria rank the five candidates in exactly reverse order, and following standard validation practice would select $\lambda = 0$, a system 4.0 times worse than the best available.

The mechanism makes the reversal systematic rather than random: a noise-corrected system deliberately *disagrees* with the erroneous labels, and C_{lr} against those labels penalises exactly that disagreement. A validation report that certifies an LR system against reconstructed reference labels therefore

measures agreement with the reconstruction, not accuracy.

J. Ablation

Table VI and Fig. 7 remove each stage in turn. Removing Stage 1 costs C_{lr}^K ($0.047 \rightarrow 0.170$ on NSL-KDD). Removing Stage 2 costs coherence ($0 \rightarrow 1.67$). Removing Stage 3 costs ALRD on NSL-KDD ($2.71 \rightarrow 5.68$) and leaves LRs of up to 26.5 decades. Replacing Stage 3 by pairwise clipping keeps C_{lr}^K but destroys coherence ($0 \rightarrow 1.41$); replacing it by scalar shrinkage keeps coherence but costs C_{lr}^K ($0.036 \rightarrow 0.111$ on CICIDS2017). Substituting the true T improves C_{lr}^K on

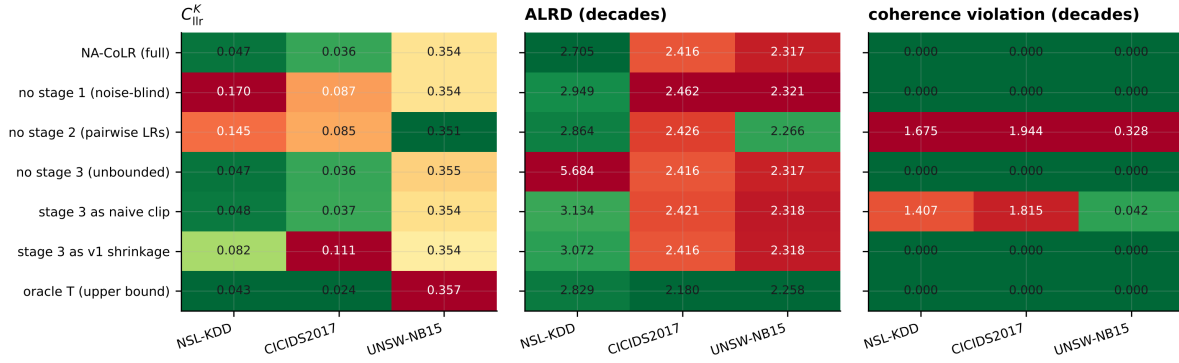


Fig. 7. Ablation heat map, column-normalised within each dataset. Green is better.

TABLE VI. ABLATION AT $\varepsilon = 15\%$

Variant	C_{lr}^K	ALRD	coh.
<i>NSL-KDD</i>			
NA-CoLR (full)	0.047	2.71	0.00
no Stage 1	0.170	2.95	0.00
no Stage 2	0.145	2.86	1.67
no Stage 3	0.047	5.68	0.00
Stage 3 as pairwise clip	0.048	3.13	1.41
Stage 3 as v1 shrinkage	0.082	3.07	0.00
oracle T	0.043	2.83	0.00
<i>CICIDS2017</i>			
NA-CoLR (full)	0.036	2.42	0.00
no Stage 1	0.087	2.46	0.00
no Stage 2	0.085	2.43	1.94
no Stage 3	0.036	2.42	0.00
Stage 3 as pairwise clip	0.037	2.42	1.81
Stage 3 as v1 shrinkage	0.111	2.42	0.00
oracle T	0.024	2.18	0.00
<i>UNSW-NB15</i>			
NA-CoLR (full)	0.354	2.32	0.00
no Stage 1	0.354	2.32	0.00
no Stage 2	0.351	2.27	0.33
no Stage 3	0.355	2.32	0.00
Stage 3 as pairwise clip	0.354	2.32	0.04
Stage 3 as v1 shrinkage	0.354	2.32	0.00
oracle T	0.357	2.26	0.00

TABLE VII. STAGE 4 AT $\varepsilon = 15\%$: C_{lr}^K POINT ESTIMATE AND 95% BOOTSTRAP BAND; TIERS FOR THE LEADING-VERSUS-RUNNER-UP PAIR (%); SHARE OF EVENTS WHOSE LEADING CLASS IS WRONG AMONG REPORTED (REP.) AND ABSTAINED (ABS.) EVENTS

	NSL-KDD	CICIDS2017	UNSW-NB15
C_{lr}^K [95% band]	0.047 [0.043, 0.078]	0.036 [0.024, 0.053]	0.354 [0.340, 0.358]
λ range	0.63–0.97	0.56–0.92	0.00–0.71
Report / Cons. / Abstain	68.0 / 30.9 / 1.2	11.1 / 88.0 / 0.9	83.4 / 0.0 / 16.6
Wrong lead: rep. / abs.	1.0 / 48.4	0.5 / 37.3	17.5 / 75.1

L. Unseen Classes

To test a change of taxonomy, each class was removed in turn from training and validation and the model was applied to a test set that still contained it. Events of the unseen class received strong evidence ($|\log_{10} \text{LR}| > 2$) for some known class in 65–99% of cases on NSL-KDD and 99–100% on CICIDS2017. A confidence gate calibrated to a 5% false-alarm rate on validation flagged only 22–70% (NSL-KDD) and 1–51% (CICIDS2017) of the unseen events. NA-CoLR therefore cannot be relied on to detect new attack types: an LR between two known classes says nothing about a third, unlisted proposition. The model must be restricted to a closed, validated taxonomy, and any addition, merger or redefinition of a class requires re-running Stages 1–4 and the validation of Section VII before reporting resumes.

VII. OPERATIONAL PROTOCOL

NA-CoLR produces evidential assessments, not conclusions. This section specifies how its diagnostics are interpreted, what is recorded, and who is accountable. Table VIII lists the decision rules.

Assumptions communicated with every report. The propositions are two classes of a closed, validated taxonomy; reference-label error is class-conditional; the LR is prior-free and must be combined with case priors only by the court; and the system has been validated on injected, not observed, reference error.

Traceability. An independent reviewer can reproduce a reported LR from the record: the hashes identify the exact inputs, the stored $\hat{T}$, λ and calibration parameters reproduce Stages 1–3 deterministically, and the seed reproduces the Stage 4 interval.

NSL-KDD and CICIDS2017 (0.043, 0.024), which locates the remaining gap in the estimation of $\hat{T}$; on UNSW-NB15 even the true T does not help (0.357).

K. Uncertainty and Abstention

Table VII summarises Stage 4. Because every replicate re-estimates $\hat{T}$, the selected λ varied over $[0.63, 0.97]$, $[0.56, 0.92]$ and $[0, 0.71]$, and this variation is now part of every interval. The C_{lr}^K bands are wider than those of a label-only resampling scheme and contain the point estimate on all three datasets, including UNSW-NB15, where label-only resampling under $\hat{T}_\lambda = I$ gives a degenerate zero-width band. For the decisive leading-versus-runner-up pair, NA-CoLR abstained on 1.2%, 0.9% and 16.6% of events. Abstention is concentrated where the attribution is wrong: the leading class was wrong for 48%, 37% and 75% of abstained events, against 1.0%, 0.5% and 17% of reported events. On CICIDS2017 most decisive pairs fall into the conservative tier (88%) because their intervals are wide, not because their direction is uncertain.

TABLE VIII. OPERATIONAL PROTOCOL: DIAGNOSTICS, DECISION RULES AND THE REPORTING RECORD

Diagnostic situation	Required action
Input check fails (Table I) $\lambda = 0$ ($\hat{T}$ unstable)	Stop. No LR is produced; the failure is logged. Report as a noise-blind system; state that reference-label error is not modelled; the technical reviewer must confirm that the validation error rate is acceptable.
$\lambda > 0$ but the bootstrap λ range exceeds 0.5 $\hat{T}$ stable but noise suspected to be instance-dependent (e.g. errors concentrated near class boundaries)	$\hat{T}$ is fragile: report the noise-blind LR alongside, flag the case, escalate to the reporting examiner. Stability does not imply correct specification (Section VI-D): compare with B3; if the reported pair differs by more than one verbal category, report the weaker value.
Corrected and noise-blind LRs differ in sign for the reported pair	Abstain for that pair; escalate.
Decisive pair in tier ABSTAIN	Report <i>inconclusive</i> ; refer the event for examination by other means (manual artefact analysis, additional sources).
Decisive pair in tier CONSERVATIVE	Report the lower bound (“at least 10^{ℓ} ”), with the interval.
Leading-class confidence below the validation gate	Possible out-of-taxonomy event: do not report an LR; escalate.
Candidate systems compared by C_{lr} on reference labels	Not admissible as a selection criterion (Section VI-I); use the corrected or clean-subset validation and record the choice.
Reporting record (written by <code>pipeline.py</code> for every case)	
Provenance	hashes of the case features and validation data; feature extraction and model versions; software versions; seed.
Model state	$\hat{T}$, λ , condition number of $\hat{T}_{\lambda}$, calibration parameters, deconvolved supports, bootstrap λ range.
Result	propositions H_i , H_j ; $\log_{10}$ LR; interval; tier; number of floored probabilities.
Assumptions and limitations	closed taxonomy; class-conditional noise; prior removal by development proportions; injected-noise validation only; any protocol rule triggered above.
Accountability	analyst, technical reviewer and reporting examiner, each with decision (accept, reject, escalate), reasons and time stamp.

Human review and accountability. Three roles are separated, in line with forensic validation guidance [1], [19]. The analyst runs the system and assembles the record but cannot release a result. The technical reviewer, independent of the analyst, checks the record against Table VIII and either accepts the assessment, rejects it, or escalates it. The reporting examiner owns the opinion and signs it; a model output is never presented as an examiner’s opinion without that signature. An examiner who disagrees with the automated output may report weaker evidence or decline to report, with written reasons in the record, but may not report stronger evidence than the system’s interval supports. Disagreements between reviewer and examiner are resolved by the laboratory’s quality manager or validation committee, and the resolution is recorded.

VIII. THREATS TO VALIDITY

1) *Injected rather than observed noise*: The reference errors are synthetic. The instance-dependent condition shows that correction can be outperformed when the class-conditional assumption fails. What is established is the mechanism and the correction under known noise models, not the magnitude in casework, which requires a corpus with observed reconstruction error and retained true labels.

2) *Revised design*: The common-floor bound and the $\hat{T}$ -resampling bootstrap were introduced during revision, after the scalar shrinkage bound and the label-only bootstrap were found deficient on the same test splits. Neither change introduces a tuned parameter, and both previous versions are reported (Tables III and VI), but the revision was informed by test-split results.

3) *Gaussian oracle*: ALRD depends on the fitted diagonal Gaussians and separates systems only on NSL-KDD.

4) *Dependence on base-model accuracy*: The method delivers nothing on UNSW-NB15, where the base model reaches 81.5% accuracy; the stability criterion detects this and disengages, but detecting a limitation is not removing it.

5) *Overstatement*: Correction raises the overstatement rate above that of the noise-blind system (0.26 against 0.18 on NSL-KDD).

6) *Software versions*: Results changed with library versions: under scikit-learn 1.9.1 and pandas 3.0.5, λ on NSL-KDD was 0.838 rather than 0.869 and the ALRD of NA-CoLR-U was 2.88 rather than 5.68 decades. Versions are therefore pinned.

7) *Domain*: The corpora are network flows. Their statistical structure (multiclass, imbalanced, tabular, reconstructed reference labels) is shared with other event-attribution problems, including database forensics, but no claim is made for database artefacts until the method is validated on them.

IX. CONCLUSION

This paper addressed three obstacles to likelihood-ratio reporting in digital forensics: calibration that assumes correct reference labels, pairwise systems whose reports contradict one another, and bounding and uncertainty machinery that is label-blind and absent, respectively. With 15% reference-label error, NA-CoLR lowered the multiclass log-LR cost relative to its noise-blind ablation by 73% and 58% on the two datasets where the transition matrix is estimable and was below the one-versus-rest baseline on both; on the third it disengaged correctly and offered no benefit. Correction helped under every noise process tested, and the lead over the one-versus-rest baseline held under pair-flip noise and prevalence shift, but not under instance-dependent noise or, on one dataset, symmetric noise. Coherence is guaranteed by construction and preserved by an order-preserving bound that costs nothing in C_{lr}^K , and intervals that propagate transition-matrix uncertainty drive an abstention policy that concentrates on erroneous attributions. Validation against erroneous labels reversed the true ranking of candidate systems, so a forensic system validated only against reconstructed reference labels has not been validated against the truth. The model does not detect previously unseen classes and must be used within a closed taxonomy and under the human-review protocol described. Validation on corpora with

observed reconstruction error, including database artefacts, is the next step.

DECLARATION ON GENERATIVE AI

The author used AI tools to support the preparation of this article, specifically to check the consistency of the references and to improve the language of the manuscript. No generative-AI tool is listed as an author. The author takes full responsibility for the accuracy, originality and integrity of the content.

DATA AND CODE AVAILABILITY

All three datasets are public and none is redistributed; the release includes a download script and the MD5 checksums of the files used. The harness (model, end-to-end pipeline with input validation and provenance record, metrics, loaders, experiment runners and figure generation) reproduces every number and figure via `./reproduce.sh` with pinned library versions, and is released under the MIT licence at <https://doi.org/10.5281/zenodo.22177199> [78]. UNSW-NB15 requires citing [73] and [74], and the Canadian Institute for Cybersecurity requires that use of NSL-KDD and CICIDS2017 cite [71] and [72]; all four are cited.

ACKNOWLEDGMENT

This research received no external funding.

REFERENCES

- [1] European Network of Forensic Science Institutes, "ENFSI guideline for evaluative reporting in forensic science," ENFSI, Tech. Rep., 2015.
- [2] National Research Council, "Strengthening forensic science in the united states: A path forward," National Academies Press, Washington, DC, Tech. Rep., 2009.
- [3] C. G. G. Aitken and F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, 2nd ed. Chichester: John Wiley & Sons, 2004.
- [4] C. Neumann, I. W. Evett, and J. Skerrett, "Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm," *Journal of the Royal Statistical Society: Series A*, vol. 175, no. 2, pp. 371–415, 2012.
- [5] G. S. Morrison, E. Enzinger, V. Hughes, M. Jessen, D. Meuwly, C. Neumann, S. Planting, W. C. Thompson, D. van der Vloed, R. J. F. Ypma, and C. Zhang, "Consensus on validation of forensic voice comparison," *Science & Justice*, vol. 61, no. 3, pp. 299–309, 2021.
- [6] H. M. R. U. Rehman, S. Liaquat, M. J. Gul, M. Z. Jhandir, D. Gavi-lanes, M. M. Vergara, and I. Ashraf, "A systematic literature study of machine learning techniques based intrusion detection: datasets, models, challenges, and future directions," *Journal of Big Data*, vol. 12, no. 1, p. 264, 2025.
- [7] A. Hozouri, A. Mirzaei, and M. Effatparvar, "A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 314, 2025.
- [8] Y. Ren, "Data science and machine learning for cyber intrusion detection: A systematic review," *Machine Learning Research*, vol. 11, no. 1, pp. 8–21, 2026.
- [9] M. Lanvin, P.-F. Gimenez, Y. Han, F. Majorczyk, L. Mé, and É. Totel, "Errors in the CICIDS2017 dataset and the significant differences in detection performances it makes," in *International Conference on Risks and Security of Internet and Systems*, 2022, pp. 18–33.
- [10] P. Goldschmidt and D. Chudá, "Network intrusion datasets: A survey, limitations, and recommendations," *Computers & Security*, vol. 156, p. 104510, 2025.
- [11] R. Zhao, J. Ding, Q. Dong *et al.*, "Mitigating label noise in network intrusion detection via graph-based sample selection and purification," *Scientific Reports*, vol. 16, p. 11674, 2026.
- [12] A. O. Prabowo, D. J. Arrizki, B. A. Pratomo, A. I. Fajar, K. B. Wijaya, H. Studiawan, A. M. Shiddiqi, and S. H. Othman, "Evaluation of anomaly-based network intrusion detection systems with unclean training data for low-rate attack detection," *Journal of Cybersecurity and Privacy*, vol. 6, no. 1, p. 14, 2026.
- [13] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2–3, pp. 230–275, 2006.
- [14] P. Vergeer, A. van Es, A. de Jongh, I. Alberink, and R. Stoel, "Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating?" *Science & Justice*, vol. 56, no. 6, pp. 482–491, 2016.
- [15] I. Ismail and K. A. Zainol Ariffin, "The admissibility of digital evidence from open-source forensic tools: Development of a framework for legal acceptance," *PLOS ONE*, vol. 20, no. 9, p. e0331683, 2025.
- [16] S. Singh and L. Devi, "Reliability and admissibility of AI-generated forensic evidence in criminal trials," *arXiv preprint arXiv:2601.06048*, 2026.
- [17] I. W. Evett and B. S. Weir, *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sunderland, MA: Sinauer Associates, 1998.
- [18] B. Robertson, G. A. Vignaux, and C. E. H. Berger, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, 2nd ed. John Wiley & Sons, 2016.
- [19] D. Meuwly, D. Ramos, and R. Haraksim, "A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation," *Forensic Science International*, vol. 276, pp. 142–153, 2017.
- [20] D. Ramos and J. Gonzalez-Rodriguez, "Reliable support: measuring calibration of likelihood ratios," *Forensic Science International*, vol. 230, no. 1–3, pp. 156–169, 2013.
- [21] M. H. DeGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *The Statistician*, vol. 32, no. 1–2, pp. 12–22, 1983.
- [22] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice*, vol. 51, no. 3, pp. 91–98, 2011.
- [23] P. Vergeer, Y. van Schaik, and M. Sjerps, "Measuring calibration of likelihood-ratio systems: A comparison of four metrics, including a new metric devpav," *Forensic Science International*, vol. 321, p. 110722, 2021.
- [24] G. S. Morrison, "Bi-Gaussianized calibration of likelihood ratios," *Law, Probability and Risk*, vol. 23, no. 1, p. mgae004, 2024.
- [25] J. Hannig and H. Iyer, "Testing for calibration discrepancy of reported likelihood ratios in forensic science," *Journal of the Royal Statistical Society: Series A*, vol. 185, no. 1, pp. 267–301, 2022.
- [26] A. Macarulla Rodriguez, Z. Geradts, and M. Worring, "Calibration of score based likelihood ratio estimation in automated forensic facial image comparison," *Forensic Science International*, vol. 334, p. 111239, 2022.
- [27] T. Guo, J. Li, and Y. Tang, "A score based likelihood ratio framework for deepfake image identification in forensic science," *Scientific Reports*, vol. 16, p. 12149, 2026.
- [28] D. Ramos, R. Haraksim, and D. Meuwly, "Likelihood ratio data to report the validation of a forensic fingerprint evaluation method," *Data in Brief*, vol. 10, pp. 75–92, 2017.
- [29] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in Neural Information Processing Systems* 26, 2013, pp. 1196–1204.
- [30] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2233–2241.
- [31] X. Xia, T. Liu, N. Wang, B. Han, C. Gong, G. Niu, and M. Sugiyama, "Are anchor points really indispensable in label-noise learning?" in *Advances in Neural Information Processing Systems* 32, 2019.
- [32] Y. Yao, T. Liu, B. Han, M. Gong, J. Deng, G. Niu, and M. Sugiyama, "Dual T: Reducing estimation error for transition matrix in label-noise learning," in *Advances in Neural Information Processing Systems* 33, 2020.

- [33] Y. Lin, R. Pi, W. Zhang, X. Xia, J. Gao, X. Zhou, T. Liu, and B. Han, "A holistic view of label noise transition matrix in deep learning and beyond," in *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.
- [34] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems 31*, 2018.
- [35] J. Li, R. Socher, and S. C. H. Hoi, "DivideMix: Learning with noisy labels as semi-supervised learning," in *International Conference on Learning Representations (ICLR)*, 2020.
- [36] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8135–8153, 2023.
- [37] N. Zhou, Q. Deng, W. Luo, X. Huang, Y. Du, B. Chen, and W. Pedrycz, "Correntropy meets cross-entropy: A robust loss against noisy labels," *Engineering Applications of Artificial Intelligence*, vol. 167, p. 113830, 2026.
- [38] R. Croft, M. A. Babar, and H. Chen, "Noisy label learning for security defects," in *Proceedings of the 19th International Conference on Mining Software Repositories (MSR)*, 2022, pp. 435–447.
- [39] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th International Conference on Machine Learning (ICML)*, 2017, pp. 1321–1330.
- [40] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd International Conference on Machine Learning (ICML)*, 2005, pp. 625–632.
- [41] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 694–699.
- [42] M. Kull, M. P. Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach, "Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration," in *Advances in Neural Information Processing Systems 32*, 2019.
- [43] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York: Springer, 2005.
- [44] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023.
- [45] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.
- [46] A. Al-Dhaqm, S. H. Othman, S. Abd Razak, and A. Ngadi, "Towards adapting metamodeling technique for database forensics investigation domain," in *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, 2014, pp. 322–327.
- [47] A. Al-Dhaqm, S. Razak, S. H. Othman, A. Ngadi, M. N. Ahmed, and A. Ali Mohammed, "Development and validation of a database forensic metamodel (DBFM)," *PLoS ONE*, vol. 12, no. 2, p. e0170793, 2017.
- [48] A. Al-Dhaqm, S. Razak, S. H. Othman, K.-K. R. Choo, W. B. Glisson, A. Ali, and M. Abrar, "CDBFIP: Common database forensic investigation processes for Internet of Things," *IEEE Access*, vol. 5, pp. 24 401–24 416, 2017.
- [49] A. Al-Dhaqm, S. Abd Razak, D. A. Dampier, K.-K. R. Choo, K. Siddique, R. A. Ikuesan, A. Alqarni, and V. R. Kebande, "Categorization and organization of database forensic investigation processes," *IEEE Access*, vol. 8, pp. 112 846–112 858, 2020.
- [50] A. Al-Dhaqm, R. A. Ikuesan, V. R. Kebande, S. Abd Razak, G. Grispos, K.-K. R. Choo, B. A. S. Al-Rimy, and A. A. Alsewari, "Digital forensics subdomains: The state of the art and future directions," *IEEE Access*, vol. 9, pp. 152 476–152 502, 2021.
- [51] A. Al-Dhaqm, W. M. S. Yafouz, S. H. Othman, and A. Ali, "Database forensics field and children crimes," in *Kids Cybersecurity Using Computational Intelligence Techniques*. Cham: Springer, 2023, pp. 81–92.
- [52] C. Grajeda, F. Breiting, and I. Baggili, "Availability of datasets for digital forensics — and what is missing," *Digital Investigation*, vol. 22, pp. S94–S105, 2017.
- [53] S. Nemetz, S. Schmitt, and F. Freiling, "A standardized corpus for SQLite database forensics," *Digital Investigation*, vol. 24, pp. S121–S130, 2018.
- [54] Z. Xu, Y. Wu, S. Wang, J. Gao, T. Qiu, and Z. Wang, "Deep learning-based intrusion detection systems: A survey," *arXiv preprint arXiv:2504.07839*, 2025.
- [55] M. M. Rahman, S. Al Shakil, and M. R. Mustakim, "A survey on intrusion detection system in IoT networks," *Cyber Security and Applications*, vol. 3, p. 100082, 2025.
- [56] A. El Mahdaoui, I. Ait Yahia, S. Oualil, and I. Berrada, "Deep learning for contextualized NetFlow-based network intrusion detection: Methods, data, evaluation and deployment," *arXiv preprint arXiv:2602.05594*, 2026.
- [57] B. Li, J. Li, and M. Jia, "ADFCNN-BiLSTM: A deep neural network based on attention and deformable convolution for network intrusion detection," *Sensors*, vol. 25, no. 5, p. 1382, 2025.
- [58] I. Koukoulis, I. Syrigos, and T. Korakis, "Self-supervised transformer-based contrastive learning for intrusion detection systems," *arXiv preprint arXiv:2505.08816*, 2025.
- [59] L. Liu, T. Cai, C. Zhou, F. Guo, K. Yao, and J. Zhou, "MEMCAIN: a memory-enhanced hybrid CNN-attention model for network anomaly detection," *Scientific Reports*, vol. 15, p. 34958, 2025.
- [60] R. Almuhanha and S. Dardouri, "A deep learning/machine learning approach for anomaly based network intrusion detection," *Frontiers in Artificial Intelligence*, vol. 8, p. 1625891, 2025.
- [61] S. W. Hall, A. Sakzad, and K.-K. R. Choo, "Explainable artificial intelligence for digital forensics," *WIREs Forensic Science*, vol. 4, no. 2, p. e1434, 2022.
- [62] A. A. Solanke, "Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models," *Forensic Science International: Digital Investigation*, vol. 42, p. 301403, 2022.
- [63] Z. Khalid, F. Iqbal, and M. Saqib, "Bridging knowledge gaps in digital forensics using unsupervised explainable AI," *Forensic Science International: Digital Investigation*, vol. 53, p. 301924, 2025.
- [64] P. Hermosilla, M. Díaz, S. Berrios, and H. Allende-Cid, "Use of explainable artificial intelligence for analyzing and explaining intrusion detection systems," *Computers*, vol. 14, no. 5, p. 160, 2025.
- [65] V. Z. Mohale and I. C. Obagbuwa, "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity," *Frontiers in Artificial Intelligence*, vol. 8, p. 1526221, 2025.
- [66] M. Chernyshev, Z. Baig, N. Syed, R. Doss, and M. Shore, "Large language models in digital forensics: capabilities, challenges and future directions," *Forensic Science International: Digital Investigation*, vol. 56, p. 302043, 2026.
- [67] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.
- [68] R. Neil and M. Zanger-Tishler, "Algorithmic bias in criminal risk assessment: The consequences of racial differences in arrest as a measure of crime," *Annual Review of Criminology*, vol. 8, pp. 97–119, 2025.
- [69] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2014.
- [70] Q. Zhang, Q. Tian, and P. Li, "Neyman–Pearson multiclass classification under label noise via empirical likelihood," *arXiv preprint arXiv:2603.21623*, 2026.
- [71] M. Tavallaei, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 1–6.
- [72] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th International Conference on Information Systems Security and Privacy (ICISSP)*, 2018, pp. 108–116.
- [73] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in *Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1–6.

- [74] —, “The evaluation of network anomaly detection systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set,” *Information Security Journal: A Global Perspective*, vol. 25, no. 1–3, pp. 18–31, 2016.
- [75] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [76] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems 30*, 2017, pp. 3146–3154.
- [77] C. G. Northcutt, A. Athalye, and J. Mueller, “Pervasive label errors in test sets destabilize machine learning benchmarks,” in *NeurIPS Datasets and Benchmarks Track*, 2021.
- [78] G. Alotibi, “NA-CoLR: A noise-aware, coherent and self-auditing likelihood-ratio engine,” 2026.