VERITY Network Intelligence Model Card v3.1.2 (External)

VERITY Network Intelligence v3.1.2

External model card · performance specification

Fidelity-based behavioral measurement for network traffic.

Credasis AI Inc. · Patent Pending · July 2026

VERITY measures whether network traffic is consistent with calibrated normal behavior. When the observed behavior and the calibrated baseline diverge, the divergence is the detection. Detection uses no attack data: thresholds derive from benign calibration only. No GPU. Air-gapped on commodity hardware under 50 MB.

Label usage: In offline evaluation, benign rows are selected using ground-truth labels (oracle benign selection). In deployment, calibration uses an assumed-benign observation window with contamination checks. Detection thresholds use only benign calibration data.

Four independent measurement perspectives operate in parallel. Each calibrates from benign only. Each captures anomalies invisible to the others. Their outputs are aggregated into a single fidelity score. Optional post-detection classification is available for labelling a detected anomaly; it plays no part in detection.

Behavioral mode discovery decomposes heterogeneous benign populations. Mode count emerges from the calibration data. Detection thresholds derive from benign calibration. The operator controls sensitivity through a single parameter (α, the false positive budget).

```
┌──────────────────────────────────────────────────────────────────────────┐
│                                                                          │
│  OPERATIONAL PROFILE                                                     │
│                                                                          │
│  Total footprint:        ~50 MB                                          │
```

```
│  Hardware requirement:   CPU only, no GPU required                    │
│  Encoding throughput:    70,000+ flows/sec (Apple Silicon)           │
│                          15,000+ flows/sec (CPU only)                │
│  Detection throughput:   1,300 to 2,100 flows/sec (full engine)      │
│  Cold-start calibration: 60 seconds on benign observation window     │
│  Recalibration:          Milliseconds (threshold re-derive)          │
│  Dependencies:           Standard scientific Python stack, CPU only  │
│  Network requirement:    None, fully air-gapped operation            │
│  Operator parameters:    One, α (false positive budget)              │
│  Tuning α:               By separation quality on calibration data   │
│  Measurement perspectives: 4, each calibrated from benign only       │
│                                                                      │
└──────────────────────────────────────────────────────────────────────┘
```

CICIDS multi-scale results independently reproduced on second hardware, agreeing to the fourth decimal in AUC.

Accuracy Profiles

All results: 5-fold cross-validated. Calibration on benign only. Default α=0.05, with a tighter operating point (α=0.005) shown alongside where relevant.

CICIDS-2017 (Enterprise, multi-scale)

Day	Flows	Attack %	AUC	±	F1 @ α=0.05	F1 @ α=0.005
Friday	500,000	46.9%	0.9990	0.000	0.972	0.992
Wednesday	496,641	35.7%	0.9986	0.000	0.957	0.995
Tuesday	322,078	2.2%	0.9985	0.000	0.470	0.898
Thursday	362,076	20.4%	0.9984	0.000	0.910	0.969

Full operating points at α=0.05: Friday P 0.946 R 0.999, Wednesday P 0.918 R 1.000, Tuesday P 0.307 R 1.000, Thursday P 0.836 R 0.997. FPR is 5.0% by construction at this α.

Tuesday shows what the α parameter does. At 2.2% attack prevalence, a 5% false positive budget admits more benign false positives than there are attacks to find, so F1 is 0.470 while AUC is 0.9985. Moving α to 0.005 recovers F1 to 0.898 with no change to the instrument.

Mode-aware calibration provides measurable improvement on heterogeneous traffic and is neutral to slightly negative where the benign population is already homogeneous. Reported as a conditional gain, not a uniform one.

CSE-CIC-IDS-2018 (Enterprise, 10 daily captures)

Random-sampled 500K flows from full daily captures (5.4M to 7.4M rows each).

Day	Flows	AUC	±	F1	Attack Types
Fri 16-02	500,000	1.000	0.000	0.948	DoS Hulk (1.8M in full capture)
Tue 20-02	500,000	1.000	0.000	0.773	DDoS-LOIC (289K in full capture)
Wed 21-02	500,000	1.000	0.000	0.929	DDoS-HOIC (1.08M in full capture)
Wed 14-02	500,000	0.993	0.000	0.781	FTP/SSH Brute Force
Thu 15-02	500,000	0.993	0.000	0.348	DoS GoldenEye/Slowloris
Fri 02-03	500,000	0.994	0.000	0.592	Botnet Ares
Wed 28-02	500,000	0.995	0.000	0.318	Infiltration
Thu 01-03	500,000	0.991	0.000	0.287	Infiltration/NMAP
Fri 23-02	500,000	0.997	0.001	0.003	Web Attack (20 flows in sample)
Thu 22-02	500,000	0.998	0.000	0.002	Web Attack (18 flows in sample)

Three AUC=1.000 days are volumetric floods, expected. Two F1≈0 days have <25 attacks in 500K, base rate. Attack-weighted F1 across all 10 days: **0.876** (versus supervised F1 0.744).

MAWIFlow (Real Backbone)

Year	Flows	Method	AUC	±	F1 @ α=0.05	F1 @ peak α
2016†	500,000	Flow-only	0.954	0.001	0.608	0.634 (α=0.07)
2021†	500,000	Multi-scale	0.924	0.001	0.521	0.547 (α=0.07)
2021†	500,000	Flow-only	0.860	0.007	0.517	0.525 (α=0.07)

† Random sample. Multi-scale uses the full measurement stack; flow-only uses the per-flow measurement alone.

Both methods are shown for 2021 because they differ materially: multi-scale ranks better (0.924 against 0.860) while flow-only reaches a similar F1 at the operating point. Each year is reported under the method that is stable for it. Multi-scale on 2016 is not stable and is not reported; multi-scale on 2021 is independently reproduced.

MAWI is the one benchmark family here drawn from real backbone traffic rather than a laboratory capture, and it is the family where AUC and the achievable operating point diverge most. Ranking on 2016 is strong at 0.954 while peak F1 is 0.634. Both are reported because the gap is the honest description of how the instrument behaves on production traffic. Do not tighten α on this traffic: at α=0.005 MAWI 2016 F1 falls to 0.150 and MAWI 2021 to 0.118.

MAWI 2011 (Real Backbone, multi-scale)

Year	Flows	AUC	±
2011	76,381	0.892	0.002

Independently reproduced at 0.8919.

Other Domains (flow-level)

Dataset	Year	Domain	Flows	AUC	±	F1	P	R	FPR
NSL-KDD Train	1999	Legacy	125,973	0.993	0.001	0.953	0.944	0.971	2.5%
UNSW Train	2015	Modern	82,332	0.993	0.002	0.961	0.960	0.977	2.3%
UNSW Test	2015	Modern	175,341	0.980	0.001	0.924	0.975	0.914	2.7%
NSL-KDD Test	1999	Legacy	22,543	0.969	0.002	0.878	0.960	0.912	2.6%
TON-IoT	2021	IoT	500,000	0.948	0.001	0.646	0.797	0.552	3.6%
DoH	2020	Encrypted	269,643	0.938	0.001	0.681	0.992	0.519	5.0%
IoT-DIAD	2024	IoT	500,000	0.864	0.003	0.504	0.779	0.691	3.0%

BCCC-IoT (2024, smart home) is withheld from this release pending end-to-end reproduction of a previously published figure.

DoH: 0.931 to 0.938 after a feature mapping correction. Precision ≥ 0.989 at all α.

GPU Cluster Behavioral Fidelity (same measurement principles)

Benchmark	Metric	Result
GWDG A100 GPU failures	Detection	16/16 (100%)
GWDG healthy baselines	False alarm	0/5 (0%)
Lingjun labeled GPU hosts	AUC (5-fold)	0.952 ± 0.007
Lingjun labeled GPU hosts	F1 at α=0.07	~0.80
Speed (Mac Mini)	Throughput	21,000+ flows/sec

Comparison to Supervised Baselines

Benchmark	VERITY (no labels)	Supervised	Method
MAWI 2016 (flow-only)	AUC 0.954	AUC 0.902	RF (Schraven et al. 2026)
MAWI 2021 (multi-scale)	AUC 0.924	AUC 0.903	RF (Schraven et al. 2026)
MAWI 2021 (flow-only)	AUC 0.860	AUC 0.903	RF (Schraven et al. 2026)
CIC-2018 (weighted F1)	F1 0.876	F1 0.744	RF (arXiv 2606.29797)
CICIDS-2017 Fri	AUC 0.9990	AUC 0.999	CNN-BiLSTM
NSL-KDD Train	AUC 0.993	AUC 0.997	XGBoost

On MAWI, label-free measurement beats the supervised baseline in 2016 by 0.052 and in 2021 by 0.021 under multi-scale. Flow-only in 2021 trails the baseline by 0.043. Both methods are shown so the margin is not overstated.

Cross-Temporal Resilience

Transfer	AUC	Notes
5-year forward (2011 → 2016)	0.942	Ranking holds; threshold must be re-derived

Transfer	AUC	Notes
10-year forward (2011 → 2021)	0.531	Comparable to supervised RF at 0.61
Same-year 2016	0.952	Cross-year feature basis
Same-year 2021	0.853	Cross-year feature basis

What transfers is the ranking, not the threshold. Thresholds must be re-derived on the benign traffic of the period being defended. That re-derivation is the ordinary recalibration path and costs minutes of benign observation.

Against the supervised comparator, ten-year transfer degrades for both approaches. The operational difference is recovery cost: recalibrate on CPU in minutes versus relabel and retrain on GPU over weeks.

Operating Point Guidance

α is the one operator parameter. The shipped default of 0.05 is not the best value on every dataset. Set α from how cleanly the instrument separates the traffic on a representative calibration sample: tighten when separation is strong, leave looser when it is weak. Tighten further when attack prevalence is low. Do not tighten on weakly separated traffic; a tight budget there starves recall.

Changing α changes no model parameter and costs nothing at runtime.

FPR Consistency

At α=0.05, observed FPR ranges 2.3% to 5.3% across all evaluated datasets (1999 to 2024). FPR is controlled by α.

Documented Limitations

Calibration assumes a clean benign observation window. Contamination checks warn the operator; automated integrity verification is still shipping.

Cross-temporal capability is ranking transfer with mandatory threshold re-derivation, not end-to-end detection transfer. Decade-scale transfer also degrades in ranking (AUC 0.531, against supervised RF 0.61).

Mode-aware calibration is a conditional gain, not a uniform one. It helps on heterogeneous benign populations and is slightly negative on homogeneous ones.

The default α understates achievable F1 on well-separated traffic. See Operating Point Guidance.

The encoder is trained on enterprise and academic benign traffic, so IoT and encrypted DNS domains are weaker (DoH AUC 0.938 after a feature mapping correction). Adversarial evasion testing against adaptive mimicry has not been conducted.

Multi-scale evaluation is not stable on every backbone capture. Where it is unstable, results are reported flow-only. Where it is stable, both methods are shown. Instability is dataset-specific rather than a general property of backbone traffic.

Credasis AI Inc. · Patent Pending

For customers, investors, and evaluators. Methodology for peer review is in the companion technical paper.