ISRG Journal of Engineering and Technology (ISRGJET)



OPEN ACCESS

ISRG PUBLISHERS

Abbreviated Key Title: ISRG J Eng Technol

ISSN: 3107-5894 (Online)

Journal homepage: https://isrgpublishers.com/isrgjet/

Volume – II Issue- IV (July – August) 2026

Frequency: Bimonthly



Post-hoc Normality Gates for Recovering Missed R2L and U2R Attacks from a Supervised Intrusion Detector: A Matched-False-Alarm Study on NSL-KDD

Khalil[1], Amir khitab[2], Abdul majid khan[3]

[1] College of computer science and software engineering, Hohai university, Nanjing, Jiangsu, China, Postal code 211100.

[2] Institute of computer science and Information technology University of Agriculture Peshawar, Pakistan

[3] School of science and technology Foundation university Rawal Pindi Campus Pakistan

| **Received:** 15-08-2026 | **Accepted:** 18-08-2026 | **Published:** 30-08-2026

***Corresponding author: Khalil**

Abstract

Supervised network intrusion detectors trained on NSL-KDD label most Remote-to-Local (R2L) attacks in the official KDDTest+ split as Normal, partly because 29.2% of test attack records belong to attack types absent from training. This paper asks how much of that loss a post-hoc gate can recover. A gate re-examines every record that an XGBoost generalist labels Normal and relabels it as the most probable attack class when an anomaly score exceeds a threshold set on held-out validation Normal records. Three scores are compared: the reconstruction error of an autoencoder trained on Normal traffic only, an Isolation Forest score, and the generalist's own confidence (one minus its normal probability). With a pre-specified 99th-percentile threshold, the autoencoder gate raised the detection rate from 68.9% to 73.9% and U2R recall from 31.3% to 43.3%, but R2L recall only from 10.5% to 11.3%, and the false alarm rate from 2.76% to 3.10% (paired bootstrap intervals exclude zero for all of these changes). Looser thresholds recovered far more R2L (50.1% at the 90th percentile) at a false alarm rate of 11.3%. The validation-derived threshold transferred poorly to the test split: across three autoencoder seeds the test false alarm rate at the 99th percentile ranged from 3.04% to 7.12%. At a matched test false alarm rate of 5%, the generalist's own confidence recovered more R2L (36.4%) than the autoencoder gate (10.6% to 25.0% across seeds), but at 10% the autoencoder recovered more (43.1% to 47.7% against 38.5%), and its ranking quality among records labelled Normal (AUC) was similar to that of the confidence score. The results show that a normal-only reconstruction gate adds some detection but does not dominate a free confidence gate, and that threshold calibration on in-distribution validation data is the main obstacle. All results are from an offline benchmark.

Keywords: *network intrusion detection; unseen attacks; autoencoder; anomaly gate; R2L; NSL-KDD; threshold calibration*

Copyright © ISRG Publishers. All rights Reserved.

DOI: 10.5281/zenodo.23941666

1. Introduction

Supervised intrusion detectors can reach accuracy above 99% on data drawn from the NSL-KDD training file, yet accuracy on the official KDDTest+ split is much lower. One reason is that 3,750 of the 12,833 attack records in KDDTest+ (29.2%) belong to attack types that never occur in KDDTrain+. A second reason is that the composition of R2L changes between the two files: 89.4% of the training R2L records are warezclient, which does not occur in the test file, whereas 75.4% of the test R2L records are guess_passwd or warezmaster. In our earlier study of a two-stage XGBoost and CNN-BiLSTM hybrid on this benchmark, 2,347 of the 2,885 R2L test records were still not labelled R2L.

A supervised model that has never seen an attack type has no reason to label it as an attack. A common answer is anomaly detection, where a model trained on Normal traffic only flags records that look unusual [7], [8]. This paper studies the simplest way to add such a model to an existing detector: a gate that re-examines only the records the supervised detector calls Normal. We ask three questions.

- RQ1. How much detection (overall and for R2L and U2R) does a normal-only autoencoder gate recover, and at what cost in false alarms?
- RQ2. Does a threshold chosen on held-out validation Normal records transfer to KDDTest+?
- RQ3. Is the autoencoder better than two cheaper scores: an Isolation Forest, and the supervised detector's own confidence?

The contributions are a controlled evaluation on the official split with a protocol that fixes thresholds on validation data only, a comparison of three gate scores at matched false alarm rates, and an analysis of how unstable the threshold transfer is across autoencoder seeds. The answer to RQ3 is less favourable to the autoencoder than we expected, and we report it as found.

2. Related Work

Three studies are closest to this work. Hindy et al. [1] trained autoencoders and one-class SVMs on NSL-KDD for zero-day attack detection and reported per-class detection rates for several reconstruction-error thresholds, with R2L detection above 95% on KDDTest+ at the thresholds they tested. Their task is binary (Normal versus attack), the threshold behaviour is reported on the test split, and no supervised multiclass detector is involved. Aygun and Yavuz [2] proposed autoencoder and denoising autoencoder models for zero-day detection and reported accuracies of 88.28% and 88.65% on KDDTest+. Again, the task is binary and the models stand alone. Finally, NetDAIL [3] combines a denoising autoencoder, an Isolation Forest score, and a LightGBM classifier for binary anomaly detection and reports an AUC of 0.998 on NSL-KDD.

These studies show that reconstruction-based scores can separate Normal from attack records on NSL-KDD. Within the three papers we read, none evaluates the autoencoder as a gate over a multiclass supervised detector, none fixes the threshold on held-out validation Normal records only and then measures how that threshold transfers to the shifted test split, and none compares the gate with the supervised detector's own confidence at matched false alarm rates. This paper addresses that gap on a deliberately small scale. We did not conduct a systematic review.

Four further lines of work frame the question. First, deep and recurrent models are the usual supervised baselines on NSL-KDD [13], [14], while the closed-world assumption behind such classifiers has long been criticised for network intrusion detection [7], and anomaly detection in general is surveyed in [8]. Second, multi-stage detectors give a second model a class-specific role: Haddad Pajouh et al. [15] used two tiers aimed at U2R and R2L, Li et al. [16] added k-nearest neighbours for uncertain connections, and Wisanwanichthan and Thammawichai [17] combined naive Bayes and a support vector machine. Third, unsupervised scorers such as Isolation Forest [11] and the autoencoder ensemble Kitsune [12] are common in network settings. Fourth, the confidence of a classifier is itself a baseline for flagging inputs it does not know: maximum softmax probability separates misclassified and out-of-distribution examples [9], although modern networks are often poorly calibrated [10]. The confidence score used in this paper follows the spirit of [9] for a gradient-boosted model.

3. Method

3.1 Data and protocol

We use NSL-KDD [4]: KDDTrain+ (125,973 records) and KDDTest+ (22,544 records), with 41 features, mapped to Normal, DoS, Probe, R2L, and U2R. KDDTrain+ is split 90/10, stratified by class with seed 42, into a fitting subset (113,375 records) and a validation subset (12,598 records). The one-hot encoder (122 features after encoding) and the min-max scaler are fitted on the fitting subset only, and negative scaled values are clipped to 0. KDDTest+ is used only for evaluation. By our count, 664 test records duplicate training records; Section 5.4 reports results without them.

3.2 Generalist

The generalist is an XGBoost classifier [18] (200 rounds, maximum depth 8, learning rate 0.1, multi:softprob, random_state 42) trained on the fitting subset after SMOTE-Tomek resampling [5], [6] (303k records after resampling). Unlike our earlier study, this run does not repair the categorical columns of synthetic rows after SMOTE, and no deep specialist is used. The generalist is trained once.

3.3 Gate

Let P be the generalist's class probabilities and s(x) an anomaly score. The gate changes only records that the generalist labels Normal:

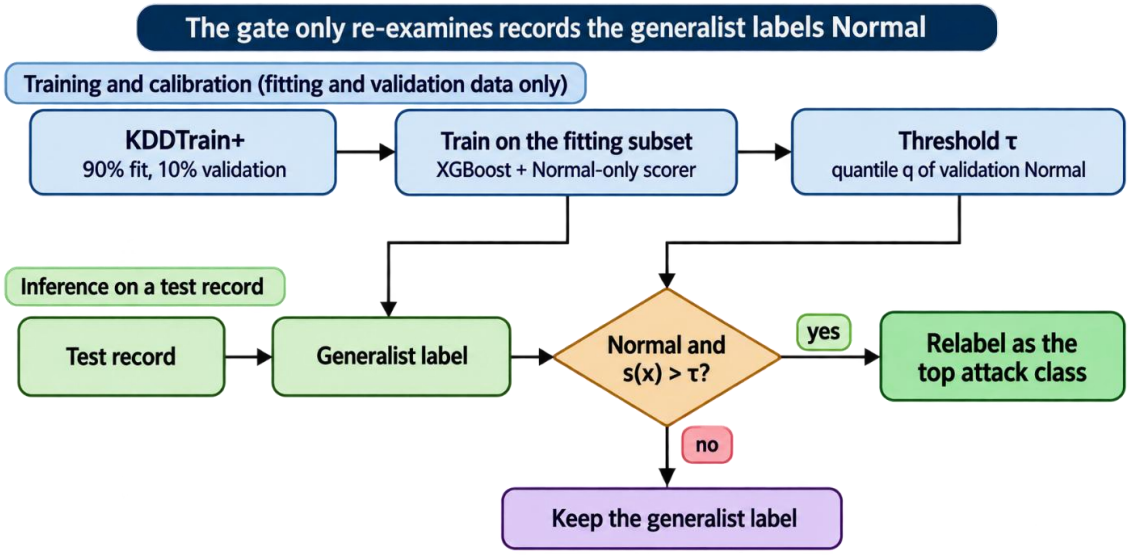
$\hat{y}$ = argmax over attack classes of P(x), if the generalist label is Normal and $s(x) > \tau$; otherwise $\hat{y}$ is the generalist label.

The relabelled class is the most probable attack class under the generalist, so the gate uses no test information. The threshold τ is the q-th quantile of s over the validation Normal records. The pre-specified setting is q = 0.99 (about 1% of validation Normal records flagged). Other values (0.98, 0.95, 0.90) are reported as a sensitivity analysis and were not used to choose q.

Three scores are compared. AE: the mean squared reconstruction error of a fully connected autoencoder (122-64-16-64-122, ReLU, Adam, batch size 256, at most 60 epochs with early stopping on a 10% internal split), trained on the 60,608 Normal records of the fitting subset only, with seeds 42, 7, and 2026. IF: the Isolation Forest anomaly score [11] (200 trees), trained on the same Normal records, with the same three seeds. CONF: one minus the

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.23941666

generalist's Normal probability, which needs no extra model and follows the confidence baseline of [9].



pipeline of the gated detector

Figure 1. Training, calibration, and inference of the gated detector. The threshold is fixed on validation Normal records before any test record is scored.

3.4 Metrics and statistics

We report accuracy, macro F1, per-class recall, the false alarm rate FAR (Normal records predicted as any attack class divided by Normal records), and the detection rate DR (attack records predicted as any attack class divided by attack records). DR counts a record as detected even if the attack family is wrong. Uncertainty for the pre-specified AE and IF gates (seed 42) uses 1,000 paired bootstrap resamples of the test records [20] with percentile 95% intervals, and McNemar's test [21] with continuity correction compares per-record correctness with the generalist. Intervals are per-metric and not adjusted for multiple comparisons. To compare scores independent of any threshold, we compute the AUC of each score among the test records the generalist labels Normal, for attack versus Normal and for R2L versus Normal. To compare gates at equal false alarm rates, we also set τ on the test scores so that FAR equals 3%, 5%, or 7%. This matched-FAR comparison uses test labels to pick τ and is therefore descriptive, not a deployable procedure.

4. Experimental Setup

Experiments ran in Python 3.x with scikit-learn 1.8.0 [19], imbalanced-learn 0.14.2, and XGBoost 3.4.1 on a single CPU core; the autoencoder is a scikit-learn MLPRegressor, so no deep learning framework was needed. The generalist alone reaches 78.75% accuracy on KDDTest+, close to the 78.20% of our earlier run with a slightly different resampling step.

5. Results

5.1 Baseline and pre-specified gates

Table 1 shows the generalist and the pre-specified (q = 0.99, seed 42) gates. The AE gate flagged 681 records, of which 146 were R2L test records and 33 were Normal records.

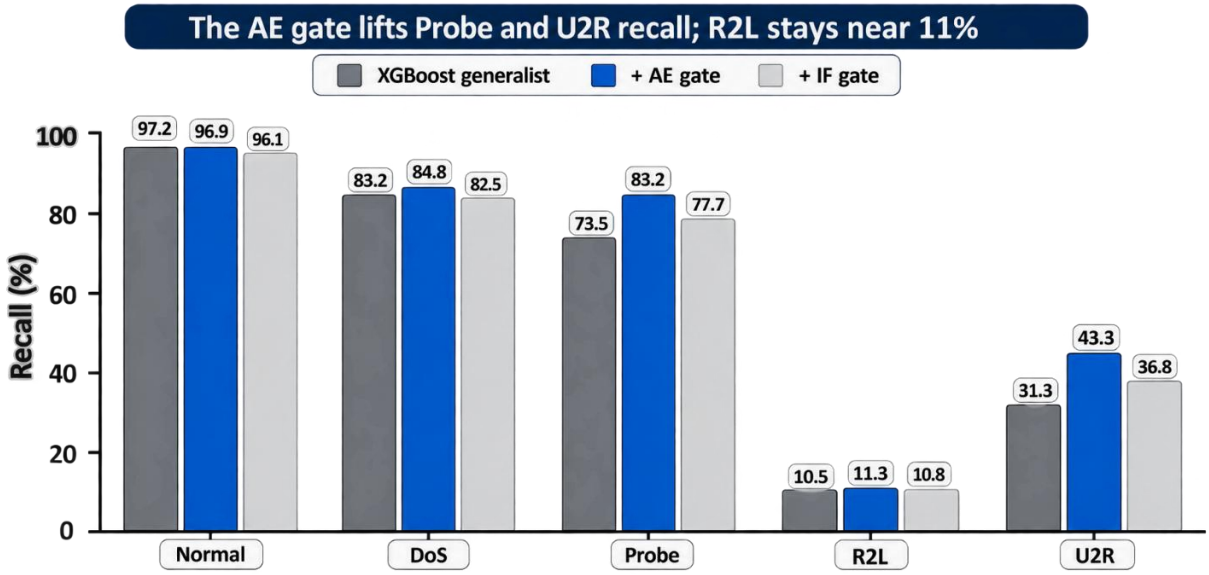
Table 1. KDDTest+ results of the generalist and the pre-specified gates (percent, except macro F1).

Model	Accuracy	Macro F1	FAR	DR	R2L recall	U2R recall
XGBoost generalist	78.75	0.6193	2.76	68.88	10.50	31.34
+ AE gate (q = 0.99)	80.33	0.6396	3.10	73.93	11.27	43.28
+ IF gate (q = 0.99)	80.12	0.6323	2.91	71.96	10.64	32.84

Paired bootstrap differences for the AE gate against the generalist (95% intervals): accuracy +1.58 points [+1.41, +1.77], macro F1 +0.0205 [+0.0040, +0.0390], R2L recall +0.76 [+0.45, +1.09], U2R recall +11.96 [+4.54, +20.59], FAR +0.34 [+0.23, +0.45], DR +5.05 [+4.68, +5.43]. McNemar's test: the gate is correct and the generalist wrong on 389 records, the reverse holds on 33 (χ^2 = 298.6, $p < 0.001$). For the IF gate: accuracy +1.38 [+1.22, +1.54], R2L +0.14 [0.00, +0.28], U2R +1.5 [0.0, +5.0], FAR +0.15 [+0.08, +0.24], DR +3.08 [+2.81, +3.39]; McNemar 325 versus 15 ($p < 0.001$).

The gain is therefore real but concentrated in DR, DoS and Probe records that the generalist labelled Normal, and U2R. R2L recall, the motivating weakness, barely changes at the pre-specified threshold.

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.23941666



per-class recall · 5 classes, 3 models, seed 42

Figure 2. Per-class recall on KDDTest+ of the generalist and the two gated detectors at the pre-specified threshold.

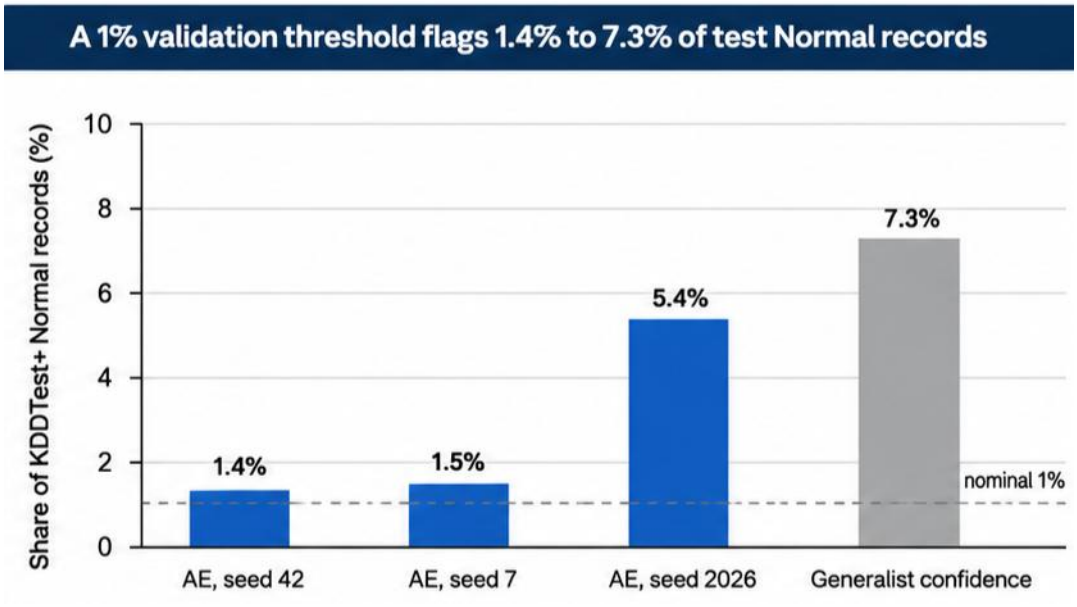
5.2 Threshold sensitivity and seed variation

Table 2 gives the mean and standard deviation over three autoencoder seeds for four validation quantiles. Lowering the threshold recovers much more R2L but costs false alarms, and the seed-to-seed spread of FAR at q = 0.99 and q = 0.98 is large.

Table 2. AE gate, mean ± SD over seeds 42, 7, and 2026 (percent).

Validation quantile q	Accuracy	FAR	DR	R2L recall	U2R recall
0.99	79.62 ± 1.15	4.42 ± 2.34	73.18 ± 0.66	11.24 ± 0.02	43.28 ± 1.50
0.98	80.72 ± 1.84	6.15 ± 2.33	77.77 ± 2.55	21.07 ± 8.43	57.21 ± 4.31
0.95	82.17 ± 0.65	9.33 ± 0.16	86.87 ± 2.41	39.04 ± 4.44	59.70 ± 0.00
0.90	82.85 ± 0.06	11.31 ± 0.08	91.84 ± 0.23	50.10 ± 0.44	61.19 ± 1.49

For comparison, the IF gate at q = 0.99 gave FAR 2.98% ± 0.06 and R2L recall 10.73% ± 0.10 over the same seeds, so it behaves almost like the generalist alone. Even at q = 0.90 the best result (R2L 50.1%) comes with a false alarm rate of 11.3%, about four times that of the generalist. A threshold of 1% on validation Normal records does not mean 1% on the test split: for the AE gate, the share of test Normal records above the q = 0.99 threshold was 1.4%, 1.5%, and 5.4% for the three seeds, whereas for the CONF score it was 7.3% (nominal 1%). The validation Normal records come from the training distribution and the test Normal records do not, so the threshold transfer is the weak point of the method (RQ2).



Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.23941666

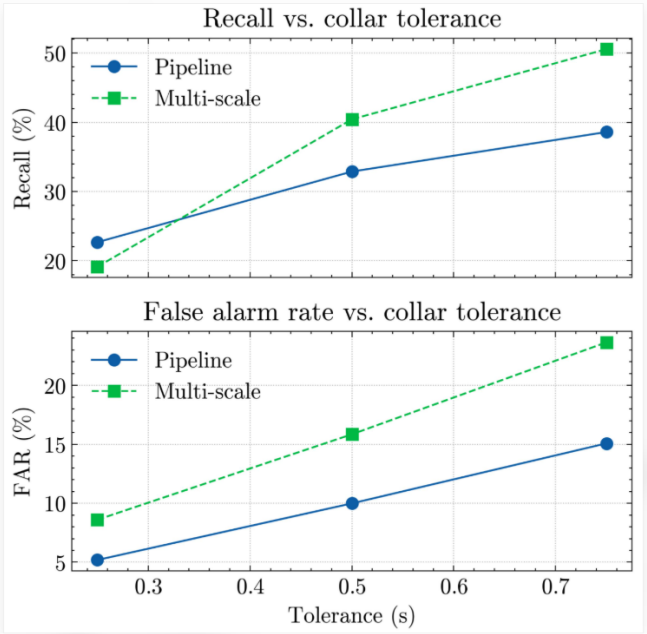
share of test Normal records flagged · 3 AE seeds and the confidence score

Figure 3. Transfer of the validation threshold to the test split. The dashed line is the nominal flag rate on validation Normal records.

5.3 Comparison of scores

Among the test records labelled Normal by the generalist, the AUC for separating attack from Normal records was 0.937, 0.930, and 0.920 for the AE score (seeds 42, 7, 2026), 0.929 for CONF. For R2L versus Normal the AUCs were 0.926, 0.921, and 0.913 for AE and 0.907 for CONF. By ranking alone the autoencoder is therefore similar to, and for two of three seeds slightly above, the generalist's own confidence.

The picture changes at fixed operating points. Figure 4 sets τ so that the gate's FAR equals 3%, 5%, 7%, 10%, or 12% on the test split (descriptive only, because test labels choose τ).



R2L recall at matched false alarm rate · 5 levels, 3 AE seeds

Figure 4. R2L recall of the confidence gate and the autoencoder gate at equal test false alarm rates.

Below about 8% the confidence gate recovers more R2L: at a 5% FAR it reaches 36.4%, against 10.6% to 25.0% for the autoencoder across seeds, with a DR of 83.7% against 69.6% to 79.3%. The order reverses at larger budgets: at 10% the autoencoder reaches 43.1% to 47.7% against 38.5%, and at 12% it reaches 51.1% to 55.5% against 43.3%. A rank-average of the two scores, calibrated on validation Normal records, gave R2L recall of 13.0% to 22.7% at q = 0.99 with a FAR of 7.4% to 8.0%, so it did not beat the confidence gate at that budget. The answer to RQ3 is therefore mixed: neither score dominates. The confidence gate is free and better at small false alarm budgets, while the autoencoder gives more R2L at budgets of 10% or more, with visible seed variation and the cost of an extra model.

5.4 Effect of duplicated test records

Without the 664 test records that duplicate training records (21,880 remain), the generalist reaches 78.33% accuracy and 0.6159 macro F1, and the AE gate 79.95% and 0.6368. The gate gain in accuracy (+1.62 points) is the same as on the full split, so the duplicates do not explain the result.

6. Discussion

What the gate does. At the pre-specified threshold the AE gate mainly converts missed DoS, Probe, and U2R records into detected ones. This is useful, but it does not solve the R2L problem that motivated the study: 11.3% R2L recall is almost unchanged. The ranking quality of the AE score is good (AUC above 0.91), so the information exists, but a 1% validation threshold cuts into it too late. Recovering R2L at a useful level needs thresholds that raise the false alarm rate to 5% to 11%.

Why thresholds fail to transfer. Validation Normal records are drawn from KDDTrain+ and score lower on average than test Normal records, because KDDTest+ contains Normal traffic with different service and flag combinations. The nominal 1% therefore turns into 1.4% to 5.4% for the AE and 7.3% for CONF. Any gate that is calibrated on in-distribution validation data inherits this shift, and in a real deployment it would be corrected only with recent labelled Normal traffic.

Practical reading. If the supervised detector already provides calibrated class probabilities, a confidence gate is a free first step and, in this benchmark, recovers more R2L than the autoencoder we trained when the false alarm budget is below about 8%. Above about 10% the autoencoder recovers more, so it may earn its extra cost where a larger false alarm budget is acceptable, for example when analysts review every flagged record.

Relation to earlier results. Hindy et al. [1] report much higher R2L detection with autoencoders on KDDTest+ in a binary setting. Our R2L numbers are lower because the gate sees only records labelled Normal by the generalist, because R2L recall requires the correct family label, and because our threshold is fixed on validation data. These studies are therefore not directly comparable with ours.

7. Limitations

- One dataset, NSL-KDD, which is old and not representative of current traffic. UNSW-NB15 and CICIDS2017 are needed.
- One generalist trained once (seed 42); only the gate models were repeated over three seeds. Part of the variation we attribute to the gate may depend on this single generalist.
- The autoencoder is a small scikit-learn network with one architecture and no hyperparameter search. Denoising or variational variants, as in the related work, may behave differently.
- Relabelling uses the most probable attack class of the generalist. Because the gate is meant to flag, DR is the fairer metric for family-agnostic detection, and family-level metrics are pessimistic.
- The matched-FAR comparison sets τ on test scores and is descriptive only. KDDTest+ was used for all secondary analyses; none was used to choose the method or the pre-specified q = 0.99, but a fresh test set is needed for confirmatory claims.
- The synthetic-row repair step and the deep specialist from our earlier study were not used, so the baseline differs slightly (78.75% versus 78.20% accuracy).

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.23941666

- Per-metric bootstrap intervals are unadjusted for multiple comparisons.

8. Conclusion

A normal-only autoencoder gate over an XGBoost generalist raised the detection rate on KDDTest+ from 68.9% to 73.9% and U2R recall from 31.3% to 43.3% at a pre-specified threshold, at a small cost in false alarms (2.76% to 3.10%), with intervals excluding zero. It did not repair R2L detection at that threshold (10.5% to 11.3%), and looser thresholds recovered more R2L only at false alarm rates of 5% to 11%. The validation-calibrated threshold transferred unreliably to the shifted test split, and at matched false alarm rates the generalist's own confidence was better below about 8% false alarms, while the autoencoder was better above about 10%. The practical message is to try a confidence gate first, to calibrate thresholds on traffic that resembles deployment, and to report operating points rather than a single threshold. Future work will repeat the study with several generalists, denoising and variational autoencoders, and a second dataset.

Declarations

Data availability: NSL-KDD is available from the Canadian Institute for Cybersecurity (https://www.unb.ca/cic/datasets/nsl.html). Code availability: [repository URL to be added]. Funding: [to be completed]. Author contributions: [to be completed]. Conflict of interest: the authors declare no conflict of interest.

References

1. H. Hindy, R. Atkinson, C. Tachtatzis, J.-N. Colin, E. Bayne, and X. Bellekens, "Utilising deep learning techniques for effective zero-day attack detection," Electronics, 2020, arXiv:2006.15344. (Verify volume, issue, and page numbers against the publisher record before submission.)
2. R. C. Aygun and A. G. Yavuz, "Network anomaly detection with stochastically improved autoencoder based models," in Proc. IEEE 4th Int. Conf. Cyber Security and Cloud Computing (CSCloud), 2017, pp. 193-198.
3. "NetDAIL: An optimized deep learning-based hybrid model for anomaly detection in network traffic," Int. J. Advanced Computer Science and Applications, vol. 16, no. 10, 2025, pp. 140-. (Add author names and final page range from the publisher page.)
4. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in Proc. IEEE Symp. Computational Intelligence for Security and Defense Applications (CISDA), 2009, pp. 1-6.
5. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.
6. G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20-29, 2004.
7. R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. IEEE Symp. Security and Privacy, 2010, pp. 305-316.
8. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, no. 3, Art. 15, 2009.
9. D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in Proc. Int. Conf. Learning Representations (ICLR), 2017.
10. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning (ICML), 2017, pp. 1321-1330.
11. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in Proc. IEEE Int. Conf. Data Mining (ICDM), 2008, pp. 413-422.
12. Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An ensemble of autoencoders for online network intrusion detection," in Proc. Network and Distributed System Security Symp. (NDSS), 2018.
13. N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," IEEE Trans. Emerging Topics in Computational Intelligence, vol. 2, no. 1, pp. 41-50, 2018.
14. C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," IEEE Access, vol. 5, pp. 21954-21961, 2017.
15. H. Haddad Pajouh, R. Javidan, R. Khayami, A. Dehghantanha, and K.-K. R. Choo, "A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks," IEEE Trans. Emerging Topics in Computing, vol. 7, no. 2, pp. 314-323, 2019.
16. L. Li, Y. Yu, S. Bai, Y. Hou, and X. Chen, "An effective two-step intrusion detection approach based on binary classification and k-NN," IEEE Access, vol. 6, pp. 12060-12073, 2018.
17. T. Wisanwanichthan and M. Thammawichai, "A double-layered hybrid approach for network intrusion detection system using combined naive Bayes and SVM," IEEE Access, vol. 9, pp. 138432-138450, 2021.
18. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794.
19. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
20. B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.
21. Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153-157, 1947.

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.23941666