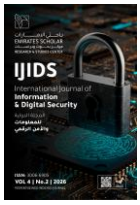

International Journal of Information & Digital Security
Vol 4 Issue 1 (2026)
Pages 23 - 31
Available at
www.emiratesscholar.com

Machine Learning for Network Intrusion Detection: An Analytical Review of Methods, Benchmark Datasets, and the Gap Between Reported and Operational Performance

Nathan Lee, Jin-Ho Park, Ethan J. Cooper
UC Berkeley - United States, Korea University - South Korea, University of Toronto - Canada

ARTICLE INFO

KEYWORDS
intrusion detection; machine learning; deep learning; base-rate fallacy; class imbalance; benchmark datasets; adversarial machine learning

HOW TO CITE
Lee, N., Park, J., & Cooper, E. J. (2026). Machine Learning for Network Intrusion Detection: An Analytical Review of Methods, Benchmark Datasets, and the Gap Between Reported and Operational Performance. International Journal of Information & Digital Security, 4(1), 23-31.
https://doi.org/10.54878/sn31gk23

ABSTRACT

Machine learning has become the dominant research approach to network intrusion detection, and the literature routinely reports detection accuracies exceeding 99% on standard benchmarks. This analytical review argues that such figures, taken at face value, are systematically misleading, and it uses the quantitative tools of detection theory to explain why. We define the confusion-matrix metrics, accuracy, precision, recall, the false-positive rate, the F1 score, and the Matthews correlation coefficient, and show that accuracy is nearly uninformative on the extreme class imbalance that characterizes real network traffic, where attacks may constitute a fraction of a percent of all flows. We then formalize the base-rate problem: because precision depends on the prior probability of attack, an intrusion detector with a 99% detection rate and a 1% false-positive rate achieves a precision below 10% when attacks are rarer than one in a hundred, so that most alarms are false. We survey the taxonomy of classical and deep-learning methods and the benchmark datasets, KDD99 and NSL-KDD, ISCX2012, UNSW-NB15, and CICIDS2017/CSE-CIC-IDS2018, on which they are evaluated, and we synthesize the reported performance, which is uniformly high. Against this we set the evidence that such performance does not transfer: benchmark datasets are dated and biased, models overfit their idiosyncrasies, concept drift erodes accuracy over time, and adversarial manipulation can defeat detectors that score near-perfectly offline. We conclude that the field's central problem is not raising benchmark accuracy, which is effectively saturated, but closing the gap between reported and operational performance through realistic evaluation, base-rate-aware metrics, robustness to drift and adversaries, and reproducible reporting. The review is methodological and is weighted toward network-based, flow-level detection.

Double-blind peer review by independent reviewers.
Received: 29 Jan 2026, Accepted: 16 May 2026, Published: 17 Jun 2026
*Corresponding author: nathan.lee@berkeley.edu

1. Introduction

The intrusion detection system, which monitors network or host activity for signs of attack, is among the oldest and most studied components of a defensive architecture, and over the past decade machine learning has become the overwhelmingly dominant approach to building one. The appeal is clear: signature-based detection catches only known attacks, whereas a model that learns the statistical difference between benign and malicious traffic promises to generalize to novel attacks as well (Buczak & Guven, 2016). The research output has been enormous, and its headline results are striking, with deep-learning models routinely reported to exceed 99% accuracy on standard benchmarks (Vinayakumar et al., 2019; Liu & Lang, 2019).

This review is written from the conviction that these headline numbers, read uncritically, do the field a disservice, and that a properly analytical reading tells a more sobering and more useful story. The gap between a detector's accuracy on a benchmark and its usefulness on a live network is not a minor caveat; it is, we argue, the central problem of the field, and it has identifiable, quantifiable causes rooted in the statistics of rare events, the limitations of the benchmark datasets, and the adversarial nature of the domain. Our purpose is to make that gap precise. We set three objectives: to define the metrics by which detection is judged and show why the most-reported of them, accuracy, is the least informative; to formalize the base-rate problem that makes high accuracy compatible with useless precision; and to survey the methods and datasets of the field and confront their impressive reported performance with the evidence that it does not transfer to operation. The analysis connects to a broader regional and applied literature on machine-learning-based threat detection, including work in this journal on intelligent models for phishing-website detection (Almalki & Abdelmajeed, 2023) and on securing cloud and Internet-of-Things systems where such detectors are increasingly deployed (Abbadi, 2025).

2. Scope and Method

This is an analytical review supported by a structured reading of the machine-learning and network-security literature. We draw on the principal surveys of the field, on the foundational papers that introduced the benchmark datasets, on representative primary studies reporting detection performance, and on the critical literature that has questioned the field's evaluation practices. Our emphasis is methodological: rather than cataloguing every proposed model, we use the standard apparatus of detection theory to interpret what reported results mean and to locate the sources of the gap between benchmark and operational performance. We concentrate on network-based, flow-level detection, where the benchmark datasets and the bulk of the literature sit, while noting that the analytical points, especially those concerning class imbalance and evaluation, apply equally to host-based and application-level detection.

3. Metrics of Detection Performance

A disciplined analysis begins with the metrics, because the field's habits of measurement are the first source of over-optimism. From the confusion matrix of true and false positives and negatives (TP, FP, TN, FN), the standard quantities are

Precision = TP / (TP + FP) , Recall (detection rate) = TP / (TP + FN) ,

FPR = FP / (FP + TN) , Accuracy = (TP + TN) / (TP + TN + FP + FN) ,

with the F1 score, the harmonic mean of precision and recall, F1 = 2·Precision·Recall / (Precision + Recall), summarizing the balance between them. The decisive observation for intrusion detection is that accuracy is dominated by the majority class. When benign traffic outnumbers attacks by a hundred or a thousand to one, a trivial classifier that labels everything benign achieves accuracy above 99% while detecting nothing, so a reported accuracy of 99% conveys almost no information about detection quality. For imbalanced problems the Matthews correlation coefficient,

MCC = (TP·TN − FP·FN) / √((TP+FP)(TP+FN)(TN+FP)(TN+FN)) ,

is a far more honest single-number summary, since it is high only when the classifier does well on both classes; yet it is reported far less often than accuracy. Table 1 collects the metrics and their properties. The analytical lesson is immediate: a study that reports only accuracy, as many do, has told the reader remarkably little, and the near-universal headline of 99%-plus accuracy on benchmarks is best read as a symptom of how the field measures rather than of how well it detects.

The choice of summary curve is subject to the same imbalance trap. The receiver operating characteristic (ROC) curve, and the area under it, plot the true-positive rate against the false-positive rate and are prevalence-independent, which is exactly why they flatter detectors on rare-event problems: a classifier can post a near-perfect ROC-AUC while its precision, and thus its practical usefulness, remains poor, because the false-positive rate axis compresses the very region, extremely

ISSN: 3006-6905 - https://doi.org/10.54878/sn31gk23

small FPR, that determines precision when positives are rare. The precision-recall curve, by contrast, exposes this region directly and is the more honest instrument for imbalanced detection. An analytically literate reading of the intrusion-detection literature therefore discounts ROC-AUC and accuracy alike and looks instead for precision at operationally realistic false-positive rates, which few papers report.

Table 1. Detection-performance metrics and their behaviour under class imbalance.

Metric	Definition	Behaviour under imbalance
Accuracy	(TP+TN)/N	Misleading; dominated by benign majority.
Recall / detection rate	TP/(TP+FN)	Fraction of attacks caught; ignores false alarms.
Precision	TP/(TP+FP)	Fraction of alarms that are real; base-rate-sensitive.
False-positive rate	FP/(FP+TN)	Small values still swamp rare positives.
F1 score	2PR/(P+R)	Balances precision and recall; prevalence-dependent.
MCC	correlation of prediction & truth	Honest single summary under imbalance.

4. The Base-Rate Problem

The deepest analytical obstacle in intrusion detection is not that detectors are inaccurate but that accuracy is the wrong thing to want, and the reason is the base-rate problem articulated for this domain a quarter-century ago and never repealed by better algorithms (Axelsson, 2000). The quantity an operator cares about is precision, the probability that an alarm is a real attack, and by Bayes' rule precision depends not only on the detector's characteristics but on the prior probability p of attack,

$$\text{Precision} = (\,DR \cdot p\,) / (\,DR \cdot p + FPR \cdot (1 - p)\,),$$

where DR is the detection rate and FPR the false-positive rate. Because malicious flows are a tiny fraction of network traffic, p is very small, and the term FPR·(1−p) in the denominator dominates unless the FPR is extraordinarily low. A worked example makes the point vivid. Consider a strong detector with DR = 0.99 and FPR = 0.01, figures that would be celebrated in most papers. If attacks constitute 1% of traffic, precision is only about 0.50, one alarm in two is false; if they constitute 0.1%, precision falls to roughly 0.09, so more than nine of every ten alarms are false; and at 0.01% it is under 0.01. Figure 2 traces this collapse for several false-positive rates. The implication is severe: at realistic base rates, even a detector with an excellent FPR generates a flood of false alarms that overwhelms analysts, the phenomenon behind much real-world alert fatigue, and reducing the FPR to the fraction of a percent required is far harder than reaching high accuracy.

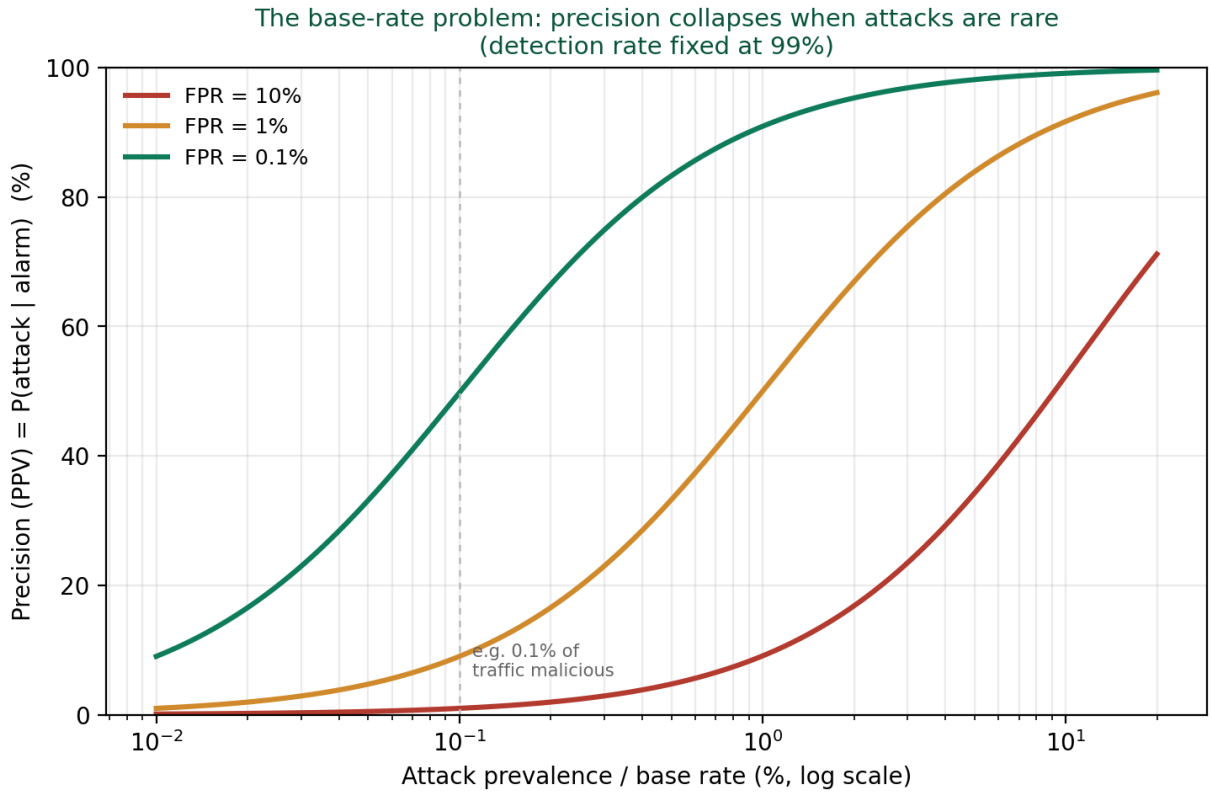

Figure 2. Precision (the probability an alarm is a genuine attack) as a function of attack prevalence, for a detector with a 99% detection rate; even a 0.1% false-positive rate yields low precision when attacks are rare.

Table 2. Precision of a detector (DR = 99%) as attack prevalence falls, for two false-positive rates.

Attack prevalence	Precision at FPR 1%	Precision at FPR 0.1%	Alarms real?
10%	0.92	0.99	mostly true
1%	0.50	0.91	half / mostly
0.1%	0.09	0.50	mostly false / half
0.01%	0.01	0.09	almost all false

The table also shows the leverage of the false-positive rate: cutting the FPR from 1% to 0.1% at a 0.1% base rate raises precision from 0.09 to 0.50, a transformation no gain in detection rate could match. This is the quantitative case for treating false-positive reduction, rather than accuracy or even detection rate, as the field's primary objective.

This single analytical fact reframes the whole enterprise. It means that the marginal value of another point of benchmark accuracy is essentially zero, while the value of an order-of-magnitude reduction in the false-positive rate at a fixed detection rate is enormous, and it means that any evaluation conducted on an artificially balanced dataset, as most benchmark evaluations are, systematically flatters the detector by hiding the base-rate penalty it will suffer in deployment. Metrics and datasets that ignore the base rate do not merely understate a difficulty; they measure the wrong thing.

5. A Taxonomy of Methods

With the evaluative caveats established, we survey the methods, organized in Figure 1. Classical machine learning remains widely used and competitive: decision trees and their ensembles (random forests, and gradient-boosting methods such as XGBoost, LightGBM, and CatBoost), support-vector machines, k-nearest neighbours, and naïve Bayes are all applied to engineered flow features, and comparative studies find that well-tuned ensembles are frequently as accurate as far more complex models while remaining interpretable and cheap to train (Hakke et al., 2025). Deep learning, the current research focus, dispenses with hand-engineered features: deep neural networks, convolutional networks applied to

traffic represented as images, recurrent and long-short-term-memory networks that model temporal structure, and autoencoders used for unsupervised anomaly detection all feature prominently, and large comparative studies report that deep models generally match or slightly exceed classical ones on the benchmarks (Vinayakumar et al., 2019; Ferrag et al., 2020). Table 2 summarizes the families. The analytically important point is that, across this whole taxonomy, reported benchmark performance is uniformly high and the differences between methods are small relative to the gap between any of them and operational reality, which suggests that method selection is no longer where the field's leverage lies.

Underlying the choice of method is a choice of representation that shapes performance at least as much. Classical models operate on features engineered from network flows, packet counts, byte volumes, inter-arrival times, protocol flags, and connection statistics, and the quality of this feature set often determines results more than the classifier does; studies repeatedly find that careful preprocessing and feature selection improve accuracy substantially, and that many benchmark gains attributed to sophisticated models in fact derive from the feature engineering around them. Deep learning promises to internalize this step, learning representations directly from raw or lightly processed traffic, which is genuinely valuable where good features are unknown, but it does not escape the representation problem so much as relocate it, since the network must still be fed inputs in a form that encodes the relevant structure, and a deep model trained on a benchmark's particular feature encoding inherits that encoding's blind spots. The analytical caution is that reported performance conflates the contributions of representation and classifier, so that a headline attributed to a new architecture may in truth belong to the preprocessing pipeline it was paired with, an ambiguity that reproducible, ablation-based reporting would resolve but that current practice usually leaves unexamined.

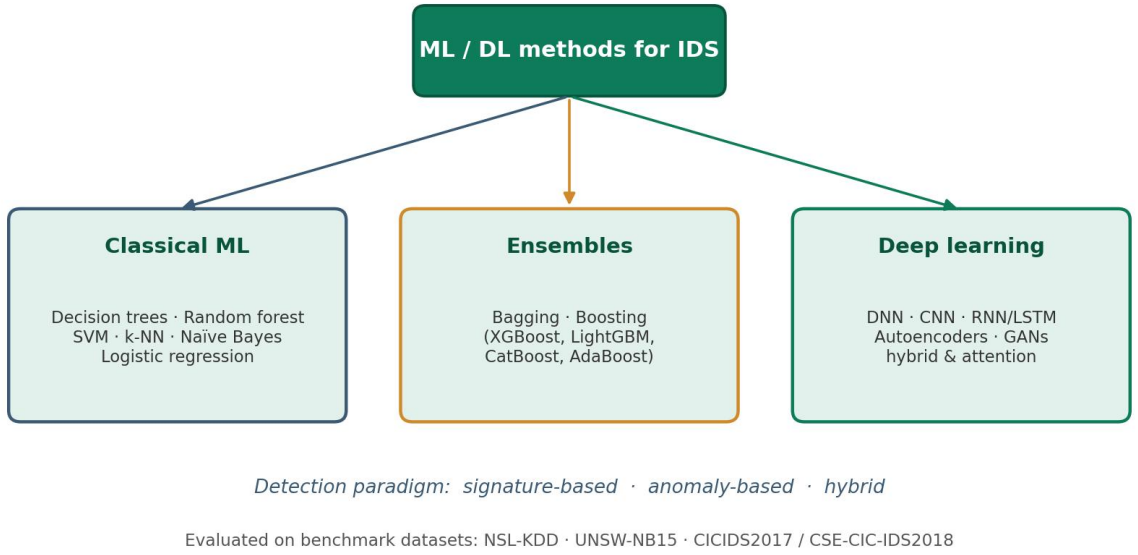

Figure 1. A taxonomy of machine-learning and deep-learning methods for intrusion detection and the paradigms and datasets on which they are evaluated.

Table 3. Principal method families for ML-based intrusion detection.

Family	Representative methods	Analytical note
Classical ML	Decision tree, SVM, k-NN, naïve Bayes.	Interpretable; competitive with tuning.
Ensembles	Random forest, XGBoost, LightGBM.	Often top benchmark accuracy; robust.
Deep learning	DNN, CNN, RNN/LSTM, autoencoders.	Automatic features; needs large data.

Unsupervised / anomaly	Autoencoders, clustering, one-class SVM.	Detects novelty; higher false-positive rate.
Adversarially aware	Robust training; detection of evasion.	Emerging; responds to evasion threat.

6. Benchmark Datasets

Because a learned detector is a function of the data that trained and tested it, the benchmark datasets deserve scrutiny in their own right, and they are, on analysis, the field's most under-acknowledged weakness. The lineage runs from the DARPA and KDD99 datasets of the late 1990s, through the cleaned NSL-KDD that removed KDD99's duplicate records (Tavallaee et al., 2009), to more modern efforts, ISCX2012 (Shiravi et al., 2012), UNSW-NB15 (Moustafa & Slay, 2015), and the CICIDS2017 and CSE-CIC-IDS2018 datasets generated from realistic testbeds with contemporary attacks (Sharafaldin et al., 2018). Detailed analyses of these datasets catalogue their instances, features, and attack classes and offer guidance on their use (Ghurab et al., 2021). Table 3 summarizes the principal ones.

Table 4. Principal benchmark datasets for network intrusion detection.

Dataset	Year	Note	Source
KDD99 / NSL-KDD	1999 / 2009	Foundational; dated attacks; NSL-KDD de-duplicated.	Tavallaee et al. (2009)
ISCX2012	2012	Profile-based realistic traffic generation.	Shiravi et al. (2012)
UNSW-NB15	2015	Modern attack families; hybrid real/synthetic.	Moustafa & Slay (2015)
CICIDS2017	2018	Realistic testbed; seven attack categories; flow features.	Sharafaldin et al. (2018)
CSE-CIC-IDS2018	2018	Large-scale cloud testbed successor.	Sharafaldin et al. (2018)

Three structural weaknesses recur. First, the datasets age quickly: attack techniques evolve, so a model trained on 2017 traffic learns a threat landscape that no longer fully exists. Second, they are imbalanced and idiosyncratic, and models can achieve high scores by exploiting artefacts of the capture, features that happen to separate classes in that dataset but carry no general meaning, rather than by learning attack behaviour. Third, and most consequentially, evaluation is almost always in-sample or cross-validated within a single dataset, so reported performance measures how well a model fits that dataset, not how well it detects intrusions in general. These are not incidental flaws; they are the mechanism by which the literature manufactures 99% accuracies that do not survive contact with a different network.

6.1 Reported performance: a quantitative snapshot

It is worth stating plainly how uniform, and how high, the reported numbers are, because their uniformity is itself diagnostic. Across the major comparative studies, deep neural networks and strong ensembles are reported to reach accuracies in the high-90s on NSL-KDD, UNSW-NB15, and CICIDS2017, with the best configurations frequently quoted above 99% and F1 scores to match (Vinayakumar et al., 2019; Ferrag et al., 2020; Hakke et al., 2025). Taken literally, such figures would imply a solved problem. The analytical reading is the opposite: when almost every method, on almost every benchmark, reports almost the same near-perfect score, the benchmark has ceased to discriminate between methods and has become a test that everything passes. A metric that no longer separates good from bad detectors has exhausted its usefulness, and continued improvement against it measures fit to the dataset rather than progress on the problem. The very saturation of benchmark performance is therefore evidence that the field's evaluation instrument, not its algorithms, is now the binding constraint.

Two corollaries follow. First, comparisons between published methods on the basis of a few tenths of a percent of benchmark accuracy are, statistically and practically, meaningless, since such differences lie within the noise of dataset splits and hyper-parameter choices

and carry no implication for operational ranking. Second, the appropriate response to a saturated benchmark is not a harder model but a harder test, one that reintroduces the difficulties, imbalance, drift, cross-network generalization, and adversarial pressure, that the benchmark removed. The field's continued optimization of a saturated metric is, in this light, a collective misallocation of effort that the analysis in this review is intended to help redirect.

7. The Generalization Gap

The consequence of the preceding analysis is a large and well-documented gap between reported and operational performance, and it has been visible for longer than the deep-learning era. A now-classic critique argued that machine learning is, for structural reasons, harder to apply to intrusion detection than to almost any other domain: the cost of errors is highly asymmetric, the base rate is punishing, benign traffic is enormously diverse so that the notion of normal is unstable, and the adversary actively adapts, all of which violate assumptions that machine learning ordinarily relies upon (Sommer & Paxson, 2010). None of these obstacles has been dissolved by newer models; they have, if anything, been obscured by benchmark scores that improved while the underlying difficulties remained.

Two further mechanisms widen the gap. The first is concept drift: network behaviour and attacks change continuously, so a model's accuracy decays after deployment unless it is retrained, and the benchmark paradigm, which trains and tests on a frozen snapshot, never measures this decay. The second is adversarial manipulation. Because an intrusion detector faces an intelligent opponent, it is subject to adversarial examples, inputs perturbed deliberately to evade detection, and a comprehensive survey shows that deep-learning detectors which score near-perfectly on clean benchmarks can be defeated by carefully crafted evasions, with the further complication that attacks and defences designed for computer vision do not transfer cleanly to the network domain (He, Kim, & Asghar, 2023). A detector's benchmark accuracy is therefore an upper bound achieved under the most favourable possible conditions, static data, no adversary, balanced classes, and the operational figure is strictly, and often substantially, lower. Recent comprehensive surveys reach the same conclusion, calling for evaluation practices that reflect these realities rather than continuing to optimize a saturated benchmark metric (Hozouri et al., 2025).

8. Challenges and the Path Forward

The analysis points to a reorientation of research priorities away from benchmark accuracy and toward the neglected determinants of operational value. Five priorities follow. First, evaluation must become base-rate-aware, reporting precision, false-positive rate, and MCC at realistic class balance rather than accuracy on balanced data. Second, models must be tested cross-dataset and, ideally, on live or streaming traffic, so that generalization rather than in-sample fit is measured. Third, robustness to concept drift must be built in and evaluated over time, through online learning and drift detection. Fourth, adversarial robustness must be treated as a first-class requirement, not an afterthought, given the intelligent adversary (He et al., 2023). Fifth, the field must adopt reproducible reporting, sharing code, data splits, and full confusion matrices, so that results can be compared and trusted. Table 4 pairs these priorities with the practices they replace.

Table 5. Reorienting evaluation: from benchmark optimization to operational validity.

Priority	Current practice	Recommended practice
Metrics	Accuracy on balanced data.	Precision, FPR, MCC at realistic base rate.
Generalization	In-sample / single-dataset CV.	Cross-dataset and live-traffic evaluation.
Concept drift	Frozen train/test snapshot.	Online learning; drift-aware evaluation.
Adversarial robustness	Clean-data testing only.	Evaluate under evasion attacks.
Reproducibility	Headline metric only.	Share code, splits, full confusion matrices.

A further operational requirement, often omitted from benchmark-driven work, is explainability. A detector that emits an alarm without a reason places the entire interpretive burden on a human analyst, who must reconstruct why the model fired before deciding whether to act; at the alarm volumes the base-rate problem

generates, this is untenable. Interpretable models, and post-hoc explanation methods for opaque ones, are therefore not a luxury but part of what makes a detector usable, and the field's drift toward ever-larger black-box deep networks sits in tension with this need. The classical methods that benchmark-chasing has tended to sideline retain an advantage here, since a decision tree or a small ensemble can often justify its output in terms an analyst can check, which in a high-false-positive regime may matter more than a marginal gain in raw detection (Hakke et al., 2025). Evaluation that ignored explainability would, once again, be measuring the wrong thing.

9. Discussion

The picture that emerges is not one of failure but of misdirected success. Machine learning genuinely can distinguish benign from malicious traffic, and on the terms the field has set itself, benchmark classification, it does so almost perfectly. The problem is that those terms measure the wrong thing. The base-rate arithmetic guarantees that a detector optimized for accuracy on balanced data will drown its operators in false alarms on a real network; the datasets guarantee that a model tuned to one of them will generalize poorly to another; and the adversary guarantees that a detector never tested against evasion will be evaded. These are analytical certainties, not empirical surprises, and they explain the enduring paradox that a field reporting near-perfect results has produced comparatively few detectors that practitioners trust in production.

It is worth being explicit that this argument is not an anti-machine-learning polemic. Learned detectors have real and durable advantages: they can surface patterns that no hand-written signature anticipates, they scale to traffic volumes no analyst could inspect, and, deployed as one layer among several rather than as an oracle, they demonstrably add value. The critique is narrower and more constructive, that the field's dominant mode of evaluation has drifted out of contact with the conditions of deployment, and that this drift, not any limitation of the underlying methods, explains the disappointment practitioners often report. Correcting the evaluation would let the genuine strengths of machine learning be directed at the parts of the problem where they help most, for instance triaging and prioritizing events for human analysts rather than autonomously adjudicating them, a role in which a moderate precision is tolerable because a human remains in the loop. Seen this way, the base-rate arithmetic is not only a warning but a design guide, indicating where in the defensive pipeline a given detector's error profile can be afforded.

The constructive implication is that the field's frontier has moved, and its metrics and datasets should move with it. The valuable research questions are no longer whether a novel architecture can add a fraction of a percent to CICIDS2017 accuracy, but how to achieve very low false-positive rates at realistic base rates, how to build detectors that generalize across networks and degrade gracefully under drift, how to withstand adversarial evasion, and how to evaluate all of this honestly and reproducibly. This reorientation is also what the applied and deployment-focused literature, including work on machine-learning detection of phishing and on securing cloud and IoT environments, ultimately requires, since those settings impose exactly the base-rate, drift, and adversarial pressures that benchmark evaluation omits (Almalki & Abdelmajeed, 2023; Abbadi, 2025). The tools of detection theory used in this review are not a counsel of despair but a map of where the real work lies.

10. Limitations

This review is analytical and interpretive rather than a fresh benchmarking study; it reasons about reported results and their evaluation rather than re-running experiments, and its illustrative computations, while grounded in the standard formulas, use representative rather than measured operating points. Its focus on network-based, flow-level detection means that its specific claims transfer less directly to host-based, application-layer, or specialized domains, though the base-rate and evaluation arguments are general. The literature is vast and fast-moving, and our selection, though centred on the most influential surveys, datasets, and critiques, reflects judgement and will date as new datasets and methods appear. These constraints qualify the detail without unsettling the central conclusion regarding the gap between reported and operational performance.

11. Conclusion

Machine learning for network intrusion detection is a field whose reported success and operational track record are strikingly out of step, and this review has used the quantitative tools of detection theory to explain why. Accuracy, the metric the field most often reports, is nearly meaningless on the extreme class imbalance of real traffic; the base-rate arithmetic ensures that even excellent detectors produce mostly false alarms when attacks are rare; the benchmark datasets are dated, biased, and evaluated in-sample so that high scores do not

generalize; and the adversarial, drifting nature of the domain further erodes performance that looks near-perfect offline. None of these is a defect of any particular algorithm; they are structural features of the problem that better classifiers cannot overcome and that benchmark optimization actively obscures. The path forward is therefore not more accuracy on saturated benchmarks but a reorientation toward operational validity: base-rate-aware metrics, cross-dataset and live evaluation, robustness to drift and adversaries, and reproducible reporting. Only measurement that reflects the real problem will produce detectors that solve it.

References

Abbadi, D. (2025). Cyber threats and risk mitigation strategies for cloud systems and the Internet of Things. International Journal of Information & Digital Security, 3(2).

Almalki, S. M., & Abdelmajeed, N. T. (2023). A new intelligent model for phishing web sites detection. International Journal of Information & Digital Security, 1(1).

Axelsson, S. (2000). The base-rate fallacy and the difficulty of intrusion detection. ACM Transactions on Information and System Security, 3(3), 186–205. https://doi.org/10.1145/357830.357849

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, 102419. https://doi.org/10.1016/j.jisa.2019.102419

Ghurab, M., Gaphari, G., Alshami, F., Alshamy, R., & Othman, S. (2021). A detailed analysis of benchmark datasets for network intrusion detection system. Asian Journal of Research in Computer Science, 7(4), 14–33. https://doi.org/10.9734/ajrcos/2021/v7i430185

Hakke, D., et al. (2025). Performance evaluation of machine learning-based intrusion detection using NSL-KDD, UNSW-NB15 and CICIDS2017 datasets. International Journal of Applied Mathematics.

He, K., Kim, D. D., & Asghar, M. R. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. IEEE Communications Surveys & Tutorials, 25(1), 538–566. https://doi.org/10.1109/COMST.2022.3233793

Hozouri, A., et al. (2025). A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. Discover Artificial Intelligence, 5, 1–35. https://doi.org/10.1007/s44163-025-00243-7

Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. Applied Sciences, 9(20), 4396. https://doi.org/10.3390/app9204396

Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS) (pp. 1–6). IEEE. https://doi.org/10.1109/MilCIS.2015.7348942

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP) (pp. 108–116). https://doi.org/10.5220/0006639801080116

Shiravi, A., Shiravi, H., Tavallaee, M., & Ghorbani, A. A. (2012). Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Computers & Security, 31(3), 357–374. https://doi.org/10.1016/j.cose.2011.12.012

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In Proceedings of the 2010 IEEE Symposium on Security and Privacy (pp. 305–316). https://doi.org/10.1109/SP.2010.25

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA) (pp. 1–6). https://doi.org/10.1109/CISDA.2009.5356528

Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. IEEE Access, 7, 41525–41550. https://doi.org/10.1109/ACCESS.2019.2895334