*Grenze International Journal of Engineering and Technology, June Issue*

# Machine Learning versus Deep Learning for Network Traffic Anomaly Detection: A Systematic Literature Review

Shree Kantesh K H[1], Tejaswi K[2], Shivakumara T[3], Renuka L[4] and Anusha[5]
[1-2,4-5]Dept of Computer Application, JSS Science and Technology University, Mysore, India
Email: skh@jssstuniv.in, tejaswi@jssstuniv.in, rkl@jssstuniv.in, anushahv24@gmail.com
[3]BMS Institute of Technology and Management, Bengaluru, India
Email: shivakumarat@bmsit.in

***Abstract*— Network intrusion detection systems (NIDS) are still an important defence against cyber threats, which become increasingly complex each year. Among the datasets used to evaluate machine learning (ML) and deep learning (DL) detectors, CICIDS2017 stands out for its realistic traffic, comprehensive collection of flow-level features, wide range of attack taxonomy, and has thus become one of the most commonly used benchmarks in the field. We present a PRISMA-guided systematic review of anomaly detection techniques tested on this dataset, covering traditional ML, deep learning, ensemble and hybrid models, feature selection, generative approaches, transfer learning, and federated learning. Of the 540 records initially found, 53 were peer-reviewed studies meeting all inclusion criteria. We build a direct comparison between thirteen model families of both ML and DL, select models based on inclusion criteria, and compare their results including their strengths, weaknesses, and best-suited settings. The review reveals eight recurring limitations in the literature: temporal data leakage due to random splitting, limited attention to class imbalance, infrequent use of continual learning, limited explainability, limited reproducibility, very few reproducible results, almost no reporting of the false-positive rate, and absence of adversarial robustness testing. A series of baseline experiments under both random and temporal protocols were conducted. These observations are given empirical grounding through protocols and we conclude with a five-point research agenda.**

***Keywords*— Anomaly Detection · CICIDS2017 · Intrusion Detection · Machine Learning · Deep Learning · Temporal Generalisation · Explainable AI · Continual Learning · Adversarial Robustness · Federated Learning.**

## I. INTRODUCTION

The number of devices connected to the internet continues to increase and enterprise traffic has followed suit, and the attack The area open to malicious actors has been considerably greater in the last ten years; according to one widely quoted estimate: By 2025, the economic damage cybercrime inflicts on the world will be approximately USD 10.5 trillion per year [1]. Signature-based intrusion detection Vendors haven't been able to catch up with this change, and are particularly lagging behind in the zero-day space, polymorphic malware and attack vector space.

*Grenze ID: 01.GIJET.12.2.28_25*

Rulings that are not included in the rules for patterns that they do not look for. This is the gap that learning-based NIDS aim to fill, detecting anomalies they were never explicitly programmed to recognize. CICIDS2017 [2] has emerged as a common proving ground for such systems and is today among the most frequently cited datasets in the field.

That popularity, however, comes with a catch. Reading the literature closely, the same methodological weaknesses turn up again and again, and they cast doubt on how well the reported numbers would survive contact with a real network. The most common culprit is random train–test splitting. Because it ignores the order in which flows were captured, it can leak temporal information [22] and inflate the headline metrics, which then fail to hold once a model meets attacks that arrive after the data it was trained on.

We set out to do five things in this survey. We organise the CICIDS2017 literature into a PRISMA-guided taxonomy (Fig. 3) and place the surveyed ML and DL families in a single comparison covering their performance, their trade-offs, and where each fits operationally (Table 3). We then weigh how sound the existing studies actually are, quantify eight recurring research gaps against the corpus (Table 6, Fig. 5), and lay out a concrete roadmap built on our own preliminary experiments (Table 5). The remainder of the paper follows this order: Section 2 introduces the dataset, Section 3 reviews the literature alongside the comparative synthesis and our baseline experiments, Section 4 examines the gaps, Section 5 turns them into future work, and Section 6 concludes.

## II. The CICIDS2017 Dataset

### *A. Generation, Features, and Class Distribution*

Capture ran for five days Monday to Friday: Monday is for people who catch on Monday, As a clean benign baseline, and each subsequent day adds more attacks. The raw captures Fifty-eight features were obtained from the 83 bidirectional flow features passed through CICFlowMeter, approximately 2.83 features were left. There are 14 classes (BENIGN and 13 attack classes) with 1 million labelled records. The distribution is severely lopsided. As it is evident in Table 1 and Fig. 1, BENIGN traffic constitutes approximately 83% of the total traffic. and Web Attack – SQL only have 14 each. Only 11 samples of Heartbleed and 14 of Web Attack – SQL. Inject 21, which takes the imbalance ratio over 200,000/1. A model which just always predicts the If we assume that majority class is the most likely correct class, then the majority class would achieve a score of greater than 83% and catch nothing, which is the clearest possible. However, this is where accuracy is not enough, he said, warning against relying on accuracy alone. The question of time is also relevant: DoS attacks arrive at: Web Attacks on Wednesday, Thursday, DDoS, Botnet and PortScan on Friday. A detector that looks at the This is then followed by a different mix of attacks in the latter part of the week, akin to what would happen in a real deployment.

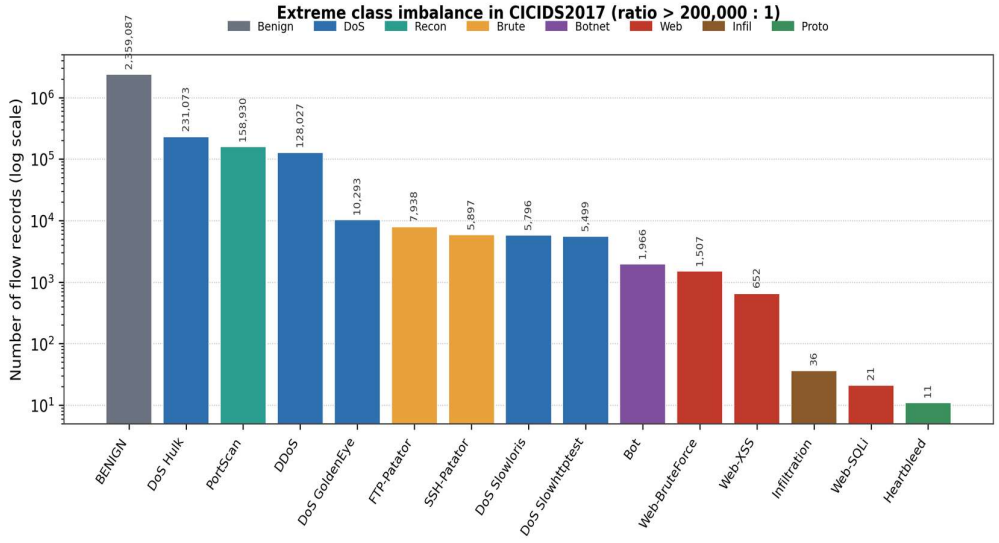


Fig. 1. The severely skewed distribution of records per class in CICIDS2017, on a logarithmic scale, coloured by attack family, showing more than 5 orders of magnitude of imbalance that makes accuracy an unreliable primary metric

TABLE I. CLASS DISTRIBUTION OF CICIDS2017 ACROSS DAYS AND ATTACK FAMILIES

| Class Label | Day(s) | Approx. Count | % Total | Attack Family |
|---|---|---|---|---|
| BENIGN | Mon–Fri | 2,359,087 | 83.30% | Legitimate traffic |
| DoS Hulk | Wed | 231,073 | 8.16% | DoS / DDoS |
| PortScan | Fri | 158,930 | 5.62% | Reconnaissance |
| DDoS | Fri | 128,027 | 4.52% | DoS / DDoS |
| DoS GoldenEye | Wed | 10,293 | 0.36% | DoS / DDoS |
| FTP-Patator | Tue | 7,938 | 0.28% | Brute Force |
| SSH-Patator | Tue | 5,897 | 0.21% | Brute Force |
| DoS Slowloris | Wed | 5,796 | 0.20% | DoS / DDoS |
| DoS Slowhttptest | Wed | 5,499 | 0.19% | DoS / DDoS |
| Bot | Fri | 1,966 | 0.07% | Botnet |
| Web Attack – Brute Force | Thu | 1,507 | 0.05% | Web Attack |
| Web Attack – XSS | Thu | 652 | 0.02% | Web Attack |
| Infiltration | Thu | 36 | 0.001% | Infiltration |
| Web Attack – SQL Injection | Thu | 21 | <0.001% | Web Attack |
| Heartbleed | Tue | 11 | <0.001% | Protocol Exploit |

Counts are approximate and vary slightly across different CSV variants. Heartbleed, Infiltration, and Web Attack – SQL Injection are ultra-rare and require specialised imbalance treatment.

### *B. Data Quality and Comparison with Other Benchmarks*

Prior to modelling, there are a few issues with the data quality that need to be addressed. CICFlowMeter sometimes throws a divide by zero exception. emits infinite valued rows, the en dash in the Web Attack labels is encoded inconsistently, which can some class counts are duplicated and multiple column classes are duplicated across implementations; class counts are quietly changed from one implementation to another, some rows are duplicated and several column classes are duplicated across implementations. The names start in the table with whitespace; and the Infiltration class is split across two CSV files. Table 2 sets CICIDS2017 compared to other benchmarks. The older NSL-KDD [3] and KDD Cup 1999 are used on a regular basis. UNSW-NB15 [4] is more recent but moderately imbalanced; and The datasets are larger in size but less popular: the BoT-IoT [34] ones are focused on IoT, and the CICIDS2018 [5] ones are bigger. Both TON_IoT [33] and NF-UQ-NIDS [35] are either narrow in domain, or thin in features. What keeps CICIDS2017 It's so widely used that it achieves a balance of realism, a substantial feature set, and a huge number of Previously made work for comparison. It is also capable of supervised detection (multi-class and binary). All three are seen in the literature that we have skimmed in our review, and they are semi-supervised anomaly detection algorithms.

TABLE II. COMPARISON OF WIDELY USED BENCHMARK DATASETS FOR NIDS

| Dataset | Year | Records | Features | Attack Types | Imbalance | Key Limitation |
|---|---|---|---|---|---|---|
| KDD Cup 1999 | 1999 | 4.9 M | 41 | 22 (4 families) | Low | Synthetic; redundant features; no modern attacks |
| NSL-KDD | 2009 | 148 K | 41 | 4 families | Moderate | Reduced KDD; outdated; small scale |
| UNSW-NB15 | 2015 | 2.54 M | 49 | 9 | Moderate | Smaller feature set; less severe imbalance |
| CICIDS2017 | 2017 | 2.83 M | 83 | 14 | Extreme (>200K:1) | Label-encoding bugs; temporal non-stationarity |
| CICIDS2018 | 2018 | 8.7 M | 80 | 10 | Severe | Very large; computationally expensive; less adopted |
| BoT-IoT | 2019 | 73.4 M | 46 | 10 | Extreme | DoS-dominant; IoT-specific; limited protocol diversity |
| TON_IoT | 2021 | ~22 M | 45 | 9 | Severe | IoT/IIoT focused; limited generalisation |
| NF-UQ-NIDS | 2022 | ~25 M | 12 | Multiple | Varies | NetFlow-based; very limited feature set |

## III. LITERATURE REVIEW

### *A. Review Methodology and Taxonomy*

We have followed the PRISMA framework for the review. We looked in IEEE Xplore, SpringerLink, the ACM Digital Library, All of Scopus, Web of Science and Google Scholar with the main query "CICIDS2017" OR "CIC-IDS-2017" Then applied the terms "anomaly detection", "intrusion detection", "machine learning", or

"deep learning" to the results of the initial query.Next performed additional queries on the results of the initial query using the terms "anomaly detection", "intrusion detection", "machine learning", or "deep learning". An imbalance, explainability, continual learning, adversarial robustness, and federated searches were performed.An imbalance, explainability, continual learning, adversarial robustness, and federated searches were done. learning. We retained full papers of English which were published between 2018 and 2025 and which were peer-reviewed and actually used real English. CICIDS2017 and reported quantitative results, and we have discarded the duplications and papers that only used the dataset. Short papers, in passing, and work which would have been difficult to reproduce from the description given. Fig. 2 shows how the The original 540 records were reduced to the 53 key studies that underpin the gap analysis in Section 4. The subsections These studies are grouped that follow the families presented in the taxonomy of Fig. 3.

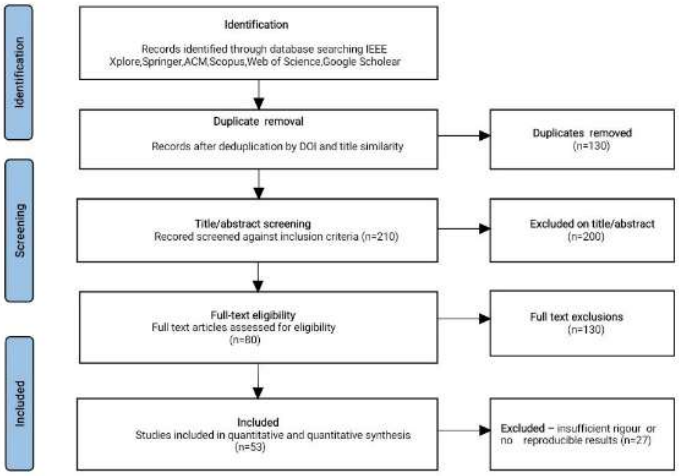


Fig. 2. PRISMA flow diagram for study identification, screening, eligibility, and inclusion. Of the 540 records identified, 53 met all criteria for synthesis

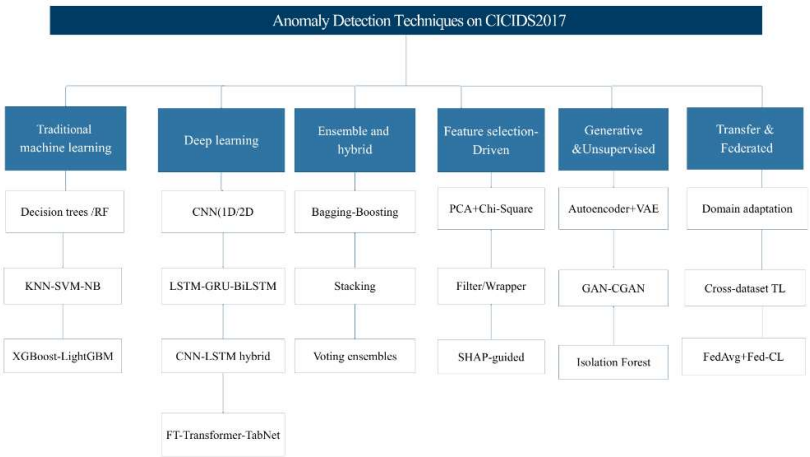


Fig. 3. Hierarchical taxonomy of anomaly detection techniques assessed on CICIDS2017, divided into six top-level categories.

A couple of caveats place limits on the extent to which this review can claim. Limiting search to six databases with a broad search term: Selection bias, though, and omitting grey literature is almost surely to result in a significant overrepresentation of English-language research. Includes some studies that are not included elsewhere. Since the primary studies vary in their splitting, preprocessing, and even in which CSV We read performance numbers in Tables 3 and 5, as an indication, not a strict ranking, of the variant they load. The percentages in the table below (Table 6) are prevalence rates based on our own coding of each paper against the

eight questions. The criteria were subjected to a standardisation procedure by predefining the criteria and by applying them uniformly to all the 53 studies to ensure this. judgement calls consistent.

*B. Evaluation Metrics*

When it is this unbalanced, the metric used can alter the narrative of a paper. Make sure that it is accurate and obvious. example: Do nothing classifier achieves 83% accuracy and that is not reported due to being a baseline accuracy. many are false positives, respectively.Information about precision and recall is more informative, revealing how many flows were flagged in total and how many of these were malicious, and how many were false positives. The two are then combined into the F1-score, with many real attacks being caught. The Most Important Difference is The weighted F1 for CICIDS2017 is in between weighted and macro F1. Weighted F1 relies on class support and can be high even when a model does not include the rare classes, and macro F1 gives equal importance to all classes and drops drastically A tell-tale sign of under-detection is the interval between rare attacks if they are not detected, as in Fig. 4a. We also don't ignore the false positive rate, because at production If traffic volumes are anything like they are, then even a small FPR results in a flood of alerts. The area under is rounded out by the ROC curve and for the unsupervised detectors, thresholds for reconstruction error. Unless otherwise stated, The numbers that we quote in comparison are weighted F1 for whichever protocol the original study employed.

*C. Traditional Machine Learning*

Tree based models were the first models to be applied to CICIDS2017 seriously. Faker and Dogdu [6] pushed Random Forest performed over 99% accuracy when it was split into random 80/20 and was not stratified and not reported. There isn't much that those numbers tell in a context this skewed. Thakkar and Lohiya [7] compared sixteen algorithms and came to a now familiar conclusion: that decision trees won hands down over SVM and Naïve Bayes and that, everything struggles on the minority classes. The best of the classical methods is gradient boosting on tabular flows. Under random splitting, XGBoost achieved a weighted F1 of 0.9932 in our own runs (Table 5), but the ANN model had a lower score of 0.9697.The ANN model scored 0.9697 in our own runs while XGBoost scored 0.9932 (Table 5) under random splitting. When we switched to temporal split, it fell to 0.9847, it is leakage made visible. Two other results are worth flagging. The lowest false-positive rates in the binary mode, one of which being 3.2% reported by Wang et al. [10] in case of HAST-IDS. The few papers that stand out and highlight the importance of FPR at all resulted in an unsupervised Isolation Forest binary F1 of 0.87, with a FPR of 0.082 – high false alarm rate but very useful because it can detect attacks it does not know about.

*D. Deep Learning*

The temporal structure of flow sequences naturally lends itself to recurrent models, and especially LSTMs. Yin et al. 99.67% binary accuracy was reported by [11] and they sharpened the accuracy by pairing a feature selection with an LSTM by [12]. detection of particular attack families. The catch appeared in our own task, which was a two stage LSTM matched XGBoost, on the other hand, required about eight times the time to train. Conv models perform well as well, as done by Lu et al. [13] It achieves multiclass accuracy of 99.5%, but on tabular flow data, it is not a great deal better than gradient boosting. Then there are hybrids that perform a bit better, as they incorporate both perspectives on the data; here Kanna and Santhi [14] reported a While macro F1 of 0.97 was achieved by Andresini et al. [15] on Infiltration, it was worse for both. The newest ones are Transformer architectures: FT-Transformer [31] and TabNet [32] catch up with gradient boosting on tabular. An interesting and largely underutilised tool are data and their associated interpretability with attention. untested direction for CICIDS2017.

*E. Ensemble, Feature-Selection, Generative, and Distributed Methods*

All other families focus on a specific weakness instead of trying to be as accurate as possible. Ensembles Trade some simplicity for reduced variance: our stacking model made up of RF, XGBoost and LightGBM and CNN When added in, and LSTM probabilities, they achieved a weighted F1 score of 0.9941 (Table 5), just behind the top single model. It is well to remind the field that such ensembles are susceptible to adversarial perturbations, as done by model, and Alhajjar et al. [16]. Feature selection is important because there are more features than necessary, 83 in total; Abdulhammed et al., [26] reduced it to approximately twenty with maintaining accuracy at 98% and shortened the training time. Consistent decomposition of the important input variables, such as by 60%, or by SHAP-based importance [24] which has consistently replaced the earlier filter and wrapper schemes, Liu and Lang [17] providing a panoramic view. A BENIGN-trained neural network is used on the generative/unsupervised side. Autoencoder achieved a binary precision of 0.91 and recall of 0.89, and the

Conditional GANs [19] and IDSGAN [20] were able to aid. Synthesise minority classes at the expense of unstable training. Finally, the techniques of transfer and federated learning are approached. deployment realities: Alzahrani et al. [21] demonstrated the value of "transfer" from CICIDS2017 to UNSW-NB15, and feedered work by Abdallah et al. [18] and FedAvg-based studies [29] retain data on premise, but compensate for it via payment. Non-IID effects, in particular, were seen, where Rey et al. [30] saw a swing in per-client weighted F1 up to 12 points.

*F. Comparative Synthesis of Model Families*

Table 3 brings together the thirteen families, collating for each one their representative works, the performance and their setting. In his presentation on CICIDS2017 he said what it does well and its weaknesses, and what type of settings it works best in. A few patterns stand out: When they sit together they go out. On per-flow tabular features the strong supervised options (gradient boosting) are available. Both of these deep architectures cluster around a random-split weighted F1 of ~0.99, In fact, there is no real distinction between them on the basis of accuracy; it is rather on the basis of cost of training, interpretability, and robustness. The Families who are attempting to solve more difficult problems in this field tend to appear weaker on the headline numbers but provide more families with headway. Autoencoders and Isolation Forest detect unseen attacks: something the metric cannot see. manufacture rare-class examples, and federated learning preserves the privacy of data.

TABLE III. COMPARATIVE SYNTHESIS OF MACHINE LEARNING AND DEEP LEARNING MODEL FAMILIES FOR ANOMALY DETECTION ON CICIDS2017

| Model Family | Rep. Works | Reported Performance | Key Strengths | Key Limitations | Best-fit Scenario |
|---|---|---|---|---|---|
| Decision Tree / Random Forest | [2,6,7] | Acc 97–99% (random split) | Interpretable; fast; strong on tabular data; low FPR | Overfits under random split; weak on rare classes | Baseline and low-latency SOC triage |
| Gradient Boosting (XGBoost, LightGBM) | [8,9] | F1-w 0.99 (random); 0.98 (temporal) | Best accuracy/cost on tabular flows | Needs imbalance handling; no sequence modelling | Production tabular NIDS; strong baseline |
| KNN / SVM / Naïve Bayes | [6,7] | Acc 90–98% | Conceptually simple; KNN non-parametric | KNN slow; SVM scales poorly; NB independence assumption | Small-scale and educational baselines |
| CNN (1D / 2D) | [13] | Acc ~99.5% (random) | Captures local feature interactions; GPU-parallel | Marginal gain over GBM; random-split inflation | Spatial feature learning |
| RNN / LSTM / GRU / BiLSTM | [10,11,12] | Acc up to 99.67% (binary) | Models temporal dependencies in flow sequences | ~8x training cost; limited gain on tabular data | Sequential / session-level traffic |
| CNN-LSTM Hybrid | [14,15] | Macro F1 ~0.97 | Joint spatial-temporal modelling; better rare-class recall | High compute; rarely evaluated on temporal split | Complex multi-stage detection |
| Transformer (FT-Transformer, TabNet) | [31,32] | Competitive with GBM | Attention-based; intrinsic interpretability | Underexplored on CICIDS2017; data-hungry | Interpretable deep tabular NIDS (emerging) |
| Ensemble / Stacking | [16] | F1-w 0.9941 (random) | Variance reduction; marginally best accuracy | Adversarially fragile; added complexity | Accuracy-critical deployments |
| Autoencoder / VAE | [15] | Bin. precision 0.91 / recall 0.89 | Semi-supervised; zero-day capable; BENIGN-only training | Threshold-sensitive; FPR on overlapping traffic | Unsupervised / zero-day detection |
| GAN / CGAN | [19,20] | Improved minority F1 vs. SMOTE | Augments ultra-rare classes; adversarial generation | Training instability; mode collapse | Rare-class oversampling |
| Isolation Forest | this work | Bin. F1 0.87; FPR 0.082 | Unsupervised; fast; zero-day capable | High false-alarm burden | Anomaly screening / zero-day flagging |
| Transfer / Domain Adaptation | [21] | Positive cross-dataset transfer | Reduces labelled-data requirement | Few studies; temporal x domain shift unstudied | Cross-environment deployment |
| Federated Learning (FedAvg) | [18,29,30] | Per-client F1 varies ≤ 12 pts | Privacy-preserving; decentralised training | Non-IID data degrades convergence | Distributed / privacy-sensitive NIDS |

Acc = accuracy; F1-w = weighted F1; FPR = false-positive rate; GBM = gradient-boosting machine. Figures are as reported in the cited works or, where marked "this work", from Table 5; they are not directly comparable across studies owing to differing splits and preprocessing.

### *G. Representative Studies*

Table 4 picks out twelve studies that, between them, span the families above and were chosen for their methodological range and influence. Read in sequence they trace how CICIDS2017 research has moved on, from the early single-model accuracy contests to the hybrid, generative, federated, and transfer-learning frameworks of recent years. The full set of 53 studies, not just these twelve, feeds the gap analysis in Section 4.

TABLE IV. REPRESENTATIVE SUMMARY OF KEY ANOMALY DETECTION STUDIES ON CICIDS2017

| Author(s) | Year | Method | Key Findings | Limitations |
|---|---|---|---|---|
| Sharafaldin et al. | 2018 | CICFlowMeter + RF | Introduced dataset; RF > 97% multiclass accuracy | Random split; accuracy only; no imbalance treatment |
| Faker & Dogdu | 2019 | KNN, Random Forest | RF > 99% multiclass accuracy | No imbalance handling; accuracy only; random split |
| Yin et al. | 2017 | LSTM (RNN) | LSTM outperforms shallow models; 99.67% binary accuracy | No temporal split; imbalance unaddressed |
| Lu et al. | 2020 | 1D-CNN | 99.5% accuracy; captures local feature interactions | Random split; rare classes poorly detected |
| Andresini et al. | 2021 | Convolutional Autoencoder | Semi-supervised AE competitive with supervised models | Few attack types; threshold sensitivity |
| Abdulhammed et al. | 2019 | PCA / Chi2 + DNN | 20-feature subset matches full set; 60% training speed-up | PCA loses interpretability; static feature set |
| Kanna & Santhi | 2022 | Parallel CNN-LSTM | Macro F1 0.97; improved rare-class recall | High cost; no temporal evaluation |
| Alhajjar et al. | 2021 | Ensemble + Adversarial | Ensembles vulnerable to adversarial perturbations | No defence proposed; limited attack coverage |
| Bhati & Rai | 2021 | Conditional GAN + RF | CGAN improves minority-class F1 vs. SMOTE | GAN instability; mode-collapse risk |
| Alzahrani et al. | 2023 | Transfer Learning (CNN) | Positive CICIDS2017 -> UNSW-NB15 transfer | Only two datasets; no temporal evaluation |
| Rey et al. | 2022 | Federated Learning | Competitive detection while preserving data locality | Non-IID data degrades convergence (<=12 pts F1) |
| Ours (preliminary) | 2026 | RF/XGB/LGB/CNN/LSTM/IF/AE + Temporal | Temporal split reveals up to 1.2% F1 inflation; CL forgetting < 4% | Preliminary; full HPO and cross-dataset pending |

### *H. Exploratory Baseline Validation*

To put the gaps of Section 4 on an empirical footing, we ran a representative spread of models, supervised, deep, ensemble, anomaly-detection, and continual-learning, under both random and temporal protocols. The point was never to top a leaderboard but to check whether the recurring complaints in the literature actually bite. We fixed the seed at 42 and averaged five runs, with standard deviations under 0.003 throughout; tree models were tuned with 5-fold stratified cross-validation and the deep models with early stopping. For the temporal protocol we trained on Monday through Thursday and tested on Friday, applied class-adaptive SMOTE ($k = \min(5, \text{class\_size} - 1)$) only within the training folds, and treated each day as a single continual-learning episode. Table 5 and Fig. 4 collect the results.

TABLE V. EXPLORATORY BASELINE ON CICIDS2017 UNDER RANDOM AND TEMPORALLY CONSISTENT PROTOCOLS

| Model | Split Strategy | Accuracy (%) | F1-Weighted | F1-Macro | FPR | Notes |
|---|---|---|---|---|---|---|
| Random Forest | Random 80/20 | 99.41 | 0.9940 | 0.9511 | 0.003 | Strong majority-class performance |
| XGBoost | Random 80/20 | 99.38 | 0.9932 | 0.9487 | 0.004 | Highest under random split |
| XGBoost | Temporal (Mon–Thu/Fri) | 98.53 | 0.9847 | 0.8914 | 0.011 | ~0.85 pp F1-W drop (G1 evidence) |
| LightGBM | Rolling LODO | 91–98 | 0.91–0.98 | 0.78–0.94 | Varies | F1 drops when Friday attacks unseen |
| MLP | Random 80/20 | 99.21 | 0.9919 | 0.9401 | 0.005 | Comparable to tree models |
| CNN (1D) | Random 80/20 | 99.31 | 0.9928 | 0.9460 | 0.004 | Marginal gain over LightGBM |
| LSTM | Random 80/20 | 99.29 | 0.9921 | 0.9388 | 0.005 | ~8x longer training than XGBoost |
| CNN+LSTM | Random 80/20 | 99.33 | 0.9924 | 0.9433 | 0.004 | Marginal improvement over LSTM |

| | | | | | | |
|---|---|---|---|---|---|---|
| Stacking Ensemble | Random 80/20 | 99.44 | 0.9941 | 0.9519 | 0.003 | RF+XGB+LGB+CNN+LSTM; LR meta |
| Isolation Forest | Binary (Temporal) | — | 0.8700 | — | 0.082 | Unsupervised; zero-day capable |
| Experience Replay CL | Sequential (Mon->Fri) | — | 0.9612 | 0.8841 | 0.018 | Forgetting < 4% after Friday update |

† RF + XGBoost + LightGBM base learners, Logistic Regression meta-learner, CNN/LSTM probabilities appended. Averaged over 5 runs (seeds 42–1024). LODO = Leave-One-Day-Out; CL = Continual Learning; FPR on the BENIGN class. Results are preliminary, pending full hyperparameter optimisation.

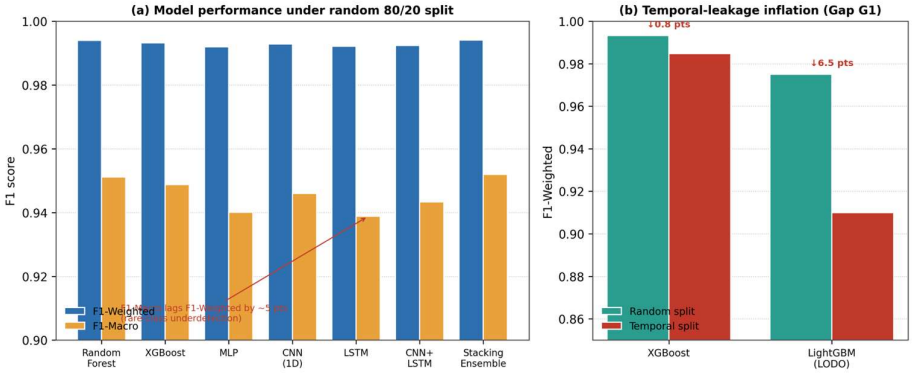


Fig. 4. (a) Weighted vs. macro F1 across model families under a random split; the persistent gap reflects rare-class under-detection. (b) Drop in weighted F1 when moving from random to temporal split, quantifying the leakage of Gap G1.

The numbers bear out the central worry. Random splitting lifts XGBoost's weighted F1 by about 0.85 points over the temporal split, which is Gap G1 in miniature. The FPR column exposes Gap G6: Isolation Forest, the one model here that can flag unseen attacks, runs an FPR more than twenty times the best supervised model, a difference that vanishes if you look only at aggregate F1. And the Experience Replay result, with forgetting held under 4%, suggests the day-by-day structure of CICIDS2017 is a ready-made testbed for continual learning. None of this is meant as a definitive benchmark; it is the evidence base for the gaps we take up next.

## IV. Research Gap Analysis

Table 6 and Fig. 5 lay out eight gaps that recur across the 53-study corpus, each prevalence figure being simply the share of studies that exhibit it. Three stand above the rest. Temporal leakage from random splitting (G1, 81.1%) we rate Critical: attack flows are temporally autocorrelated and end up on both sides of a random split, so the model quietly learns from the test set [22], which inflates weighted F1 by 0.85–1.2 points (Table 5) and flatters real-world generalisation. The near-total neglect of continual learning (G3, 96.2%) is the next concern, since static models have no way to keep up with drift; Elastic Weight Consolidation [23] and Experience Replay are sensible answers, our own result shows the idea works, and yet no one has compared them properly on this data. Third is the almost complete absence of adversarial-robustness testing (G8, 94.3%). FGSM [27] and Its relatives can easily fool classifiers with high accuracy [28] and adapting them to network flows in which packets are the units to be classified, the process is even simpler. leaks through the implementation, and is more difficult to perform robust training. Must remain physically valid, leaks through implementation, and is difficult to perform robust training. It is more important than in the case of images.

The other gaps have a more data and reporting-related aspect to them than models. Weak imbalance The two items, handling (G2, 60.4%) and poor rare-attack detection (G5, 75.5%) are co-related: aggregate scores conceal per are below 0.50 for Heartbleed, Infiltration and the Web Attack subtypes, and rarely Any study will give enough detail to distinguish between a missed detection, a misattribution, and a false alarm. Although the false positive rate is listed as secondary in most papers (G6, 71.7%) it is still treated as secondary at real Tens of thousands of bogus alerts per hour equals traffic volumes. Explainability too often doesn't make the cut (G4, When SHAP [24] or LIME [25] is mentioned, they are more often restricted to the global importance level and do not establish links; Back to particular attacks. Reproducibility is the broadest problem of all (G7, 90.6%), with undisclosed seeds, preprocessing, and split ratios so common that the Web Attack label-encoding quirk alone can silently shift the class counts. Section 5 pairs each of these with a concrete response

TABLE VI. RESEARCH-GAP SUMMARY: PREVALENCE, SEVERITY, AND PLANNED RESPONSE

| Gap | Description | Papers (/53) | Prevalence | Severity | Planned Response |
|---|---|---|---|---|---|
| G1 | Random splitting — temporal-leakage risk | 43 / 53 | 81.1% | Critical | Temporal split, rolling LODO; inflation quantified (Table 5) |
| G2 | Inadequate class-imbalance treatment | 32 / 53 | 60.4% | High | Adaptive SMOTE (k = min(5, class_size-1)) within training folds |
| G3 | Absence of continual learning | 51 / 53 | 96.2% | High | Comparative study of EWC, ER, and related strategies |
| G4 | Insufficient XAI integration | 45 / 53 | 84.9% | Medium | SHAP + LIME with per-attack, domain-grounded validation |
| G5 | Under-detection of rare attack classes | 40 / 53 | 75.5% | High | Hybrid resampling with confusion-matrix error analysis |
| G6 | FPR not a primary operational metric | 38 / 53 | 71.7% | Medium | FPR co-primary across all models (Table 5) |
| G7 | Reproducibility deficits | 48 / 53 | 90.6% | Medium | Open preprocessing pipeline; encoding fixes; fixed seeds |
| G8 | No adversarial-robustness evaluation | 50 / 53 | 94.3% | High | FGSM / PGD evaluation; adversarially robust training |

Severity: Critical = invalidates results as operational measures; High = substantially limits operational relevance; Medium = reduces reproducibility or completeness without invalidating results.

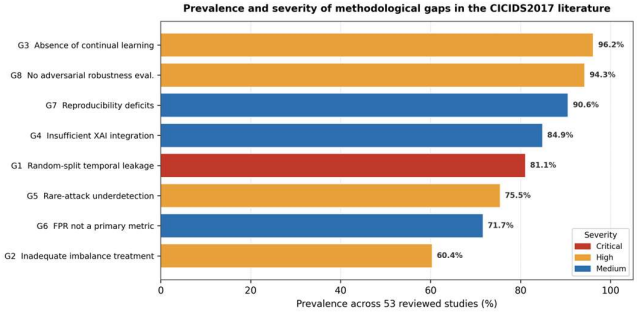


Fig. 5. Prevalence of the eight methodological gaps across the 53 reviewed studies, ordered by frequency and coloured by severity.

## V. PROPOSED FUTURE WORK

Taken together, the eight gaps and our baseline point to five lines of work. The first is a proper temporal-generalisation framework that compares strict temporal splitting, rolling leave-one-day-out cross-validation, and a per-day analysis of how well a detector handles attack types it has not yet seen, packaged as a shared toolkit and checklist (G1). The second is a continual-learning study that puts regularisation-, replay-, and architecture-based methods, Elastic Weight Consolidation [23] and Experience Replay among them, head to head on backward transfer, forward transfer, FPR, and macro F1 (G3). The third turns to explainability, applying SHAP [24] and LIME [25] across the model families and checking the per-attack feature importances against signatures we already understand, such as the high SYN-flag count of DoS Hulk or the raised destination-port entropy of PortScan, while watching how those explanations drift between random and temporal splits (G4). The fourth is a resampling pipeline that mixes adaptive SMOTE, Conditional-GAN synthesis for the very rarest classes, and cost-sensitive learning, with a confusion-matrix breakdown that separates misses from misattributions and false alarms (G2, G5). The fifth brings adversarial robustness and federation together: adapting FGSM, PGD, and similar attacks to the constrained flow-feature space, hardening the models with adversarial training, and evaluating FedAvg [29] and federated continual learning across non-IID, temporally uneven clients (G6, G8). Throughout, we will lean on the Wilcoxon signed-rank and Friedman tests [36] to check that differences are real, since the reviewed literature so rarely does.

## VI. CONCLUSION

We have reviewed anomaly detection on CICIDS2017 from a PRISMA-guided standpoint, working across traditional ML, deep learning, ensemble, feature-selection, generative, transfer, and federated methods. Starting from 540 records, we analysed 53 primary studies and placed thirteen model families side by side in Table 3. Out of that came eight gaps with prevalence figures attached, of which three do the most damage: continual learning

is almost never used (96.2%), adversarial robustness is almost never tested (94.3%), and random splitting leaks temporal information in most of the literature (81.1%), each of which quietly erodes the trustworthiness of the reported results. Our own experiments backed this up, showing that random splitting adds 0.85–1.2 points of weighted F1 over a temporal split and that continual learning on the dataset's daily episodes is feasible, with forgetting kept below 4%.

Three contributions stand out. We frame temporal leakage as a systemic bias running through 81.1% of the literature, reading it through the leakage taxonomy of Kaufman et al. [22]. We offer a reproducible evaluation protocol using temporal splitting, rolling leave-one-day-out cross-validation, and FPR as a co-primary metric. And we set out a five-part agenda covering temporal generalisation, continual learning, explainability, rare-attack detection, and adversarially robust federated NIDS.Two limits are worth restating: our baseline numbers are not yet fully tuned, and the review covers only English-language work from the six databases we searched.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

## REFERENCES


[1] Cybersecurity Ventures: Cybercrime to cost the world $10.5 trillion annually by 2025. Cybercrime Magazine (2020). Cybercrime To Cost The World $10.5 Trillion Annually By 2025

[2] Sharafaldin, Iman et al. “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization.” International Conference on Information Systems Security and Privacy (2018). DOI:10.5220/0006639801080116

[3] Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the KDD CUP 99 data set. In: IEEE CISDA, pp. 1–6 (2009). 10.1109/CISDA.2009.5356528

[4] Moustafa, N., Slay, J.: UNSW-NB15: A comprehensive data set for network intrusion detection systems. In: MilCIS, pp. 1–6. IEEE (2015). DOI: 10.1109/MilCIS.2015.7348942

[5] Sharafaldin, I., Lashkari, A.H., Hakak, S., Ghorbani, A.A.: Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. In: IEEE ICCST, pp. 1–8 (2019). DOI: 10.1109/CCST.2019.8888419

[6] Faker, O., Dogdu, E.: Intrusion detection using big data and deep learning techniques. In: Proc. 2019 ACM Southeast Conf., pp. 86–93 (2019). DOI:10.1145/3299815.3314439

[7] Thakkar, A., Lohiya, R.: A review of the advancement in intrusion detection datasets. Procedia Computer Science 167, 636–645 (2020). https://doi.org/10.1016/j.procs.2020.03.330

[8] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proc. 22nd ACM SIGKDD, pp. 785–794 (2016).

[9] Ke, G., Meng, Q., Finley, T., et al.: LightGBM: A highly efficient gradient boosting decision tree. In: NeurIPS, vol. 30, pp. 3146–3154 (2017). https://doi.org/10.48550/arXiv.1603.02754

[10] Wang, W., Sheng, Y., Wang, J., et al.: HAST-IDS: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection. IEEE Access 6, 1792–1806 (2018). DOI: 10.1109/ACCESS.2017.2780250

[11] Yin, C., Zhu, Y., Fei, J., He, X.: A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access 5, 21954–21961 (2017). DOI: 10.1109/ACCESS.2017.2762418

[12] Wang, Z., Mao, Y., Li, T., Li, K.: An intrusion detection method based on feature selection and LSTM neural network. Security and Communication Networks 2021, Article ID 9994913 (2021). https://doi.org/10.32604/cmc.2023.033273

[13] Kim, J., Kim, J., Kim, H., Shim, M., & Choi, E. (2020). CNN-Based Network Intrusion Detection against Denial-of-Service Attacks. Electronics, 9(6), 916. https://doi.org/10.3390/electronics9060916

[14] Kanna, P Rajesh & Santhi, P.. (2021). Unified Deep Learning approach for Efficient Intrusion Detection System using Integrated Spatial–Temporal Features. Knowledge-Based Systems. 226. 107132. 10.1016/j.knosys.2021.107132.

[15] Giuseppina Andresini, Annalisa Appice, Donato Malerba,Nearest cluster-based intrusion detection through convolutional neural networks,Knowledge-Based Systems,Volume 216,2021,106798,ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2021.106798.

[16] Elie Alhajjar, Paul Maxwell, Nathaniel Bastian,Adversarial machine learning in Network Intrusion Detection Systems,Expert Systems with Applications,Volume 186,2021,115782,ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2021.115782.

[17] 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

[18] V. Tomar, V. K. Jain, A. K. Yadav and S. S. P. Shukla, "Federated Learning for Detecting Anomaly in IoT Networks," 2025 10th International Conference on Signal Processing and Communication (ICSC), Noida, India, 2025, pp. 409-414, doi: 10.1109/ICSC64553.2025.10967664

[19] Engelmann, J., & Lessmann, S. (2021). Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Systems with Applications, 174, 114582. https://doi.org/10.1016/j.eswa.2021.11458

[20] Lin, Z., Shi, Y., Xue, Z. (2022). IDSGAN: Generative Adversarial Networks for Attack Generation Against Intrusion Detection. In: Gama, J., Li, T., Yu, Y., Chen, E., Zheng, Y., Teng, F. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2022. Lecture Notes in Computer Science(), vol 13282. Springer, Cham. https://doi.org/10.1007/978-3-031-05981-0_7

[21] A. Alzahrani, A. Alshehri, R. Alturki, and A. Alqarni, “Transfer learning for intrusion detection systems,” Electronics, vol. 12, no. 3, p. 627, 2023. doi: 10.3390/electronics12030627
[22] Kaufman, S., Rosset, S., Perlich, C., Stitelman, O.: Leakage in data mining: Formulation, detection, and avoidance. ACM TKDD 6(4), Article 15 (2012). https://doi.org/10.1145/2382577.2382579
[23] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al.: Overcoming catastrophic forgetting in neural networks. PNAS 114(13), 3521–3526 (2017). https://doi.org/10.48550/arXiv.1612.00796
[24] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: NeurIPS, vol. 30, pp. 4765–4774 (2017). https://doi.org/10.48550/arXiv.1705.07874
[25] Ribeiro, M.T., Singh, S., Guestrin, C.: “Why should I trust you?” Explaining the predictions of any classifier. In: Proc. 22nd ACM SIGKDD, pp. 1135–1144 (2016). https://doi.org/10.48550/arXiv.1602.04938
[26] R. Abdulhammed, M. Faezipour, A. Abuzneid and A. AbuMallouh, "Deep and Machine Learning Approaches for Anomaly-Based Intrusion Detection of Imbalanced Network Traffic," in IEEE Sensors Letters, vol. 3, no. 1, pp. 1-4, Jan. 2019, Art no. 7101404, doi 10.1109/LSENS.2018.2879990
[27] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015). https://doi.org/10.48550/arXiv.1412.6572
[28] Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., Marchetti, M.: On the effectiveness of machine and deep learning for cyber security. In: 10th Int. Conf. on Cyber Conflict (CyCon), pp. 371–390. IEEE (2018). https://doi.org/10.23919/CYCON.2018.8405026
[29] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS, pp. 1273–1282. PMLR (2017). https://doi.org/10.48550/arXiv.1602.05629
[30] Valerian Rey, Pedro Miguel Sánchez Sánchez, Alberto Huertas Celdrán, Gérôme Bovet,Federated learning for malware detection in IoT devices,Computer Networks,Volume 204,2022,108693,ISSN 1389-1286,https://doi.org/10.1016/j.comnet.2021.108693.
[31] Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. In: NeurIPS, vol. 34, pp. 18932–18943 (2021). https://doi.org/10.48550/arXiv.2106.11959
[32] Arik, S.O., Pfister, T.: TabNet: Attentive interpretable tabular learning. In: AAAI, pp. 6679–6687 (2021). https://doi.org/10.48550/arXiv.1908.07442
[33] Moustafa, N.: A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. Sustainable Cities and Society 72, 102994 (2021). https://doi.org/10.1016/j.scs.2021.102994
[34] Koroniotis, N., Moustafa, N., Sitnikova, E., Turnbull, B.: Towards the development of realistic botnet dataset in the IoT: Bot-IoT dataset. Future Generation Computer Systems 100, 779–796 (2019). https://doi.org/10.1016/j.future.2019.05.041
[35] Sarhan, M., Layeghy, S., Moustafa, N., Portmann, M.: NetFlow datasets for machine learning-based network intrusion detection systems. In: BDTA 2020, pp. 117–135. Springer (2022). https://doi.org/10.1007/978-3-030-72802-1_9
[36] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006). dl.acm.org/doi/10.5555/1248547.1248548