JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

Machine Learning Techniques for TCP/IP-Based Intrusion Detection Systems: A Systematic Literature Review

Mukaila Ishola Adediran*, S. A. Abdullahi, Salma Muhammad Abidi, Kenneth Y. Giwa

Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria

*Corresponding Author: mikkysho26@gmail.com

ABSTRACT

The evidence base for Machine Learning-based intrusion detection is fragmented across incompatible benchmarks and unvalidated deployment claims. This PRISMA 2020 compliant systematic review synthesised 50 primary studies (2017 - 2026) from five electronic databases across four research questions on technique taxonomy, benchmark adequacy, comparative performance, and open challenges. Classical ML (random forest dominant, n = 23) and deep learning (LSTM/CNN dominant, n = 24) are co-dominant paradigms, ensemble methods yield at most 0.003 F1 macro gain over their strongest base classifier with no reliable diversity advantage on the principal benchmarks. Evaluation concentrates on three legacy datasets (NSL-KDD, CICIDS-2017, UNSW-NB15) with documented validity limitations with only three studies use the NF-v2 NetFlow family, the sole framework enabling controlled cross study comparison. Temporal interarrival features raise multiclass macro F1 by 19.8-27.3 percentage points over static flow features while binary gains remain below 2.2 points, exposing a systematic masking effect in aggregate accuracy reporting. Minority class failure is mechanistically rooted in feature poverty rather than sample scarcity. Eight convergent gaps are identified: adversarial robustness is untested in all 50 studies; concept drift is addressed in only one. Structural reform of the evaluation paradigm is required before ML-based IDS research can produce reproducible, comparable, and deployable results.

KEYWORDS: Intrusion Detection System, Machine Learning, Deep Learning, TCP/IP Networks, Systematic Literature Review, Network Security, Prisma 2020, Benchmark Evaluation, Minority Class Detection.

I. INTRODUCTION

Network intrusion detection systems (NIDS) are a critical component of defence-in-depth security architectures, re-sponsible for identifying malicious traffic and anomalous behaviour in real time (Eltanbouly et al., 2025). The increasing volume, velocity, and diversity of network attacks have made rule-based and signature-based IDS insufficient for contemporary threat environments as such systems cannot generalise to attack variants that fall outside their predefined signature databases (Abdulboriy and Shin, 2024). Machine learning (ML) has been extensively applied to intrusion detection over the past decade as a means of learning discriminative traffic representations directly from data without requiring manually engineered rules (Le Jeune et al., 2021). The TCP/IP protocol stack is the substrate of virtually all modern networked systems, attack detection at the network and transport layers provides the earliest point of intervention before application layer damage occurs (Luay et al., 2026). Flow level features derived from TCP/IP packet captures (source and destination addresses, port numbers, protocol flags, packet sizes and inter arrival times) constitute the primary input representation for ML-based IDS and the choice of features fundamentally determines which attack types a model can and cannot detect (Luay et al., 2026). Current machine learning-based intrusion detection systems (IDS) frequently struggle to detect rare attack types, partly because researchers evaluate them using outdated benchmark datasets that lack modern TCP/IP traffic patterns. Furthermore, studies seldom validate these systems under real-world deployment conditions (Eltanbouly et al., 2025; Francis et al., 2026).

Several surveys of ML-based IDS have been published. Earlier reviews catalogued technique families and reported performance across benchmark datasets (Aldweesh et al., 2020; Khraisat et al., 2019).

https://doi.org/10.5281/zenodo.23385324

More recent work has examined dataset quality and labelling errors (Eltanbouly et al., 2025), feature selection methods for IoT IDS (Francis et al., 2026) and streaming detection under operational constraints (Ahmed et al., 2020). None of these reviews applied PRISMA 2020 methodology, conditioned performance comparisons on shared evaluation protocols, or identified a mechanistic explanation for the persistent failure of minority class detection. The benchmark standardisation gap and the latency reporting gap have been identified individually in prior surveys (Eltanbouly et al., 2025; Francis et al., 2026) but not quantified across a common primary study corpus under shared evaluation conditions. The objective of this review is to systematically identify, appraise and synthesise published evidence on machine learning techniques applied to TCP/IP based network intrusion detection to characterise technique families and their detection performance to evaluate benchmark adequacy and to identify methodological gaps that constrain reproducibility and deployment readiness. This paper presents a PRISMA 2020 compliant systematic literature review of machine learning techniques for TCP/IP based network intrusion detection. Fifty primary studies published between 2017 and 2026 were identified through searches of five electronic databases and synthesised around four research questions:

RQ1: What machine learning technique families have been applied to TCP/IP based IDS and what detection performance do they achieve?

RQ2: What datasets are used for evaluation and how adequately do they represent real world TCP/IP traffic?

RQ3: How do technique families compare in detection performance when evaluation conditions are held constant?

RQ4: What are the principal methodological gaps and deployment barriers in the current body of work?

The remainder of this paper is structured as follows: Section 2 describes the search and selection methodology. Section 3 presents synthesis findings organised by research question and Section 4 discusses implications, limitations, and directions for future research.

II. METHODOLOGY

This study adopted a systematic literature review design, conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta Analyses 2020 statement (Page et al., 2021). The systematic review methodology was selected because it provides a reproducible, transparent, and bias minimising framework for synthesising evidence across a heterogeneous body of primary studies (Page et al., 2021). No meta-analysis was performed as the included studies were too heterogeneous in their evaluation protocols, dataset choices and reported metrics to permit quantitative pooling. The review was guided by the following formal question: In TCP/IP based network environments, what machine learning techniques have been applied to intrusion detection and what evidence exists regarding their detection performance, benchmark adequacy and deployment readiness? The review was guided by four research questions. RQ1 addressed the taxonomy of machine learning technique families applied to TCP/IP based intrusion detection and the performance each family achieves across reported metrics. RQ2 examined the datasets used for evaluation and the extent to which they represent real world TCP/IP traffic. RQ3 compared detection performance across technique families, comparative claims were conditioned on evaluation conditions rather than raw accuracy figures which are unreliable under class imbalance. RQ4 identified methodological gaps deployment barriers and under investigated problem areas in the body of work. The structuring of RQ2 around benchmark fitness rather than simple dataset enumeration distinguishes this review from prior surveys on the topic most of which catalogue datasets without interrogating their suitability for TCP/IP specific deployment.

A. Eligibility Criteria

Eligibility criteria were specified in the review protocol prior to any database search in accordance with PRISMA 2020 guidance (Page et al., 2021). The criteria were structured using the Population, Intervention, Comparison and Outcome (PICO) framework summarised in Table 1.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

Table 1. PICO framework for the systematic review.

Element	Definition
Population	TCP/IP based network environments including enterprise local area networks, cloud hosted networks, Internet of Things networks and industrial control systems where intrusion detection operates at the network or transport layer of the TCP/IP stack.
Intervention	Machine learning techniques applied to intrusion detection, including classical ML (support vector machines, decision trees, random forests, *k*-nearest neighbours, naive Bayes), ensemble methods (boosting, bagging, stacking) and deep learning (convolutional neural networks, recurrent neural networks, autoencoders, transformers), operating on TCP/IP traffic features.
Comparison	Alternative ML approaches, rule-based or signature-based IDS, hybrid systems, or no explicit comparator. All were accepted at screening.
Outcome	*Primary:* detection performance measured by accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve, or false positive rate. *Secondary:* computational overhead, model size, inference latency, dataset representativeness and real-world deplorability.

The specific inclusion and exclusion criteria derived from this framework are presented in Table 2. Three borderline decision rules were prespecified to govern the most probable screening ambiguities. Studies proposing a hybrid system that combined signature based and ML-based detection were included only when ML constituted a primary detection component with its own reported performance metrics, studies in which ML served solely as a post filter with no isolated evaluation were excluded. Studies conducted in Internet of Things or industrial control system environments were included when detection operated on TCP/IP traffic features derived from packet captures or network flow records, studies relying exclusively on application layer payloads with no TCP/IP feature extraction were excluded. Studies using synthetic or simulated benchmark datasets were included when the dataset was an established, publicly available benchmark, studies using privately generated datasets with no reproducibility information and no public release were excluded and flagged separately in the screening record.

Table 2: Inclusion and Exclusion Criteria

Criterion	Inclusion	Exclusion
Population	TCP/IP IDS: packet, flow, or session-level analysis.	Application layer only security (SQLi, XSS); no TCP/IP modelling.
Intervention	Any ML/DL technique applied to IDS (classical, ensemble, neural).	Rule-based, signature-based, or threshold methods only; no ML.
Comparison	Any comparator or none.	-
Outcome	≥1 quantitative detection metric (accuracy, F1, precision, recall, AUC, FPR).	Purely conceptual; no experimental evaluation.
Study design	Peer reviewed journal articles and conference proceedings.	Reviews, surveys, meta analyses, editorials, book chapters.
Language	English	Non-English
Date range	2017-2026.	Before 2017
Publication type	Peer reviewed journals; indexed conference proceedings.	Preprints, technical reports, theses, grey literature.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

B. Search Strategy

Five electronic databases were searched: IEEE Xplore, SpringerLink, ScienceDirect, Taylor & Francis Online and PubMed. All searches were conducted in July 2026, a publication date filter of 2017 to 2026 was applied in each database interface. Only English language publications were retrieved. The databases were selected on four grounds. First, disciplinary fit: IEEE Xplore, SpringerLink and ScienceDirect host the principal journals and indexed conference proceedings in computer networking, cybersecurity and applied machine learning, and the venues that recur in the included corpus (for example IEEE Access, Scientific Reports, Knowledge-Based Systems and Journal of Big Data) are published on these platforms; 38 of the 50 included studies carry IEEE, Springer Nature or Elsevier DOI prefixes. Second, consistency with the eligibility criteria: the selected sources are peer-reviewed publisher platforms, which supports the exclusion of preprints, technical reports and grey literature in Table 2 and allows peer-review status to be verified at screening. Third, reproducibility: all five interfaces support structured Boolean queries, date filtering and bulk export of records with DOIs, which made the search strings repeatable and enabled DOI-based deduplication before screening. Fourth, breadth: Taylor & Francis Online was added to extend coverage to applied computing and security journals outside the three largest publisher platforms, and PubMed was added to capture interdisciplinary work in healthcare network security and medical Internet of Things environments. Using five sources rather than one reduces dependence on the indexing practices of any single platform, although it does not make the search exhaustive.

Three databases that would otherwise have been candidates for inclusion (ACM Digital Library, Scopus, and Web of Science) were unavailable to the research team. The consequences for comprehensiveness are specific. The ACM Digital Library indexes security and networking conference proceedings, so its absence mainly risks missing conference-only studies. Scopus and Web of Science are multi-publisher citation indexes, so their absence mainly risks missing studies in venues outside the five searched platforms and removes the opportunity to cross-check recall against an independent index. Because the searched sources are largely publisher platforms rather than multi-publisher indexes, the retrieved corpus is best interpreted as a systematic sample of the literature available within the searched sources and not as a complete census of machine learning-based intrusion detection research. This constraint is carried into the interpretation of results in Section 4.3. Search strings were constructed by mapping the PICO elements to three Boolean blocks. Block A covered population terms: TCP/IP, network traffic, network intrusion, computer network and network security. Block B covered intervention terms both classical and deep learning technique names were included: machine learning, deep learning, neural network, random forest, support vector machine, convolutional neural network, LSTM, XGBoost, autoencoder and ensemble learning. Block C covered application terms: intrusion detection, anomaly detection, attack detection, NIDS and network intrusion detection system. The three blocks were combined with AND terms within each block were combined with OR. Outcome terms such as accuracy, F1 score and area under the receiver operating characteristic curve were deliberately excluded from the search strings and applied only at the screening stage to avoid inadvertently excluding studies that report secondary metrics only.

Database specific interface constraints required minor adaptations to this structure. IEEE Xplore was searched using the Command Search interface with all metadata field tags applied to each term. SpringerLink required two subqueries because its interface does not support field prefixes reliably, one subquery covered core ML terms and a second covered architecture specific variant. ScienceDirect imposed a limit of eight Boolean operators per query, necessitating four subqueries partitioned by technique family. Taylor & Francis Online was searched with a single string across the default title, abstract and keywords scope export records from this database were found to contain erroneous year metadata and were verified against source pages before screening. PubMed was searched using Title/Abstract field tags. This returned a smaller but complementary set that captured interdisciplinary work in healthcare network security and medical Internet of Things environments. The searches returned records across all five databases prior to deduplication, full counts are reported in Section 3. A proportion of export records contained no abstract, concentrated in SpringerLink and PubMed exports where metadata completeness varies by publisher agreement. Abstract absent records were screened on title

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

alone where the title provided insufficient information, the DOI was used to retrieve the abstract directly from the source page. This gap in export metadata is acknowledged as a limitation of the search process.

C. Study Selection

Study selection proceeded in two stages; all authors participated in screening and disagreements were resolved by discussion until consensus was reached.

Stage 1: Title and abstract screening. All deduplicated records from the five database exports were assessed independently by all authors against the eligibility criteria in Table 2. Records absent an abstract were screened on title alone; DOI retrieval was used where the title was ambiguous. Disagreements were resolved by discussion where consensus was not reached and the majority view was adopted.

Stage 2: Full-text assessment. Records passing Stage 1 were retrieved in full and assessed against all eight eligibility criteria and the three borderline decision rules in Section 2.2. Reasons for exclusion were recorded for each rejected article and are reported in the PRISMA flow diagram (Section 3). Conflict resolution followed the same procedure as Stage 1. The PRISMA flow diagram documenting record counts at each selection stage is presented in Section 3.

D. Data Extraction

A structured extraction form was applied uniformly to all included studies; fields were specified in the review protocol prior to extraction. Table 7 (Appendix B) lists the fields collected from each paper. Extraction was performed by all authors working independently on each record, discrepancies in extracted values were resolved by returning to the source document and reaching consensus. No automated extraction tools were used; all fields were populated manually from the full text of each included study.

E. Quality Assessment

Each included study was assigned an overall quality rating using a three-tier scheme developed for this review. The scheme was applied to the primary detection methodology and evaluation design of each paper; secondary contributions were noted but did not determine the rating. Table 8 (Appendix C) defines the criteria for each tier.

The tiers were operationalised as an ordered decision rule. A study was rated Strong only if it met all five Strong conditions: three or more independent datasets, re-implemented or directly comparable baselines, per-class F1 reported for minority attack classes, at least one efficiency indicator (latency, model size or throughput), and no irreconcilable methodological inconsistencies. A study was rated Adequate if its design was sound and its primary claims were supported by the reported evidence, with at most one of five prespecified limitations: a single evaluation dataset, published-figure baselines only, minority class performance unreported, efficiency unmeasured, or statistical significance untested. A study was rated Weak if it showed at least one of the disqualifying problems listed in Table 8, such as irreconcilable dataset and results inconsistencies or no quantitative baseline on the same dataset. The ratings describe how well a study's evidence supports its own claims; they were not used as an exclusion criterion, and all 50 studies were retained in the synthesis.

Quality ratings were assigned independently by all authors and reconciled by discussion until consensus was reached. No chance-corrected inter-reviewer agreement statistic (for example Cohen's or Fleiss' kappa) was calculated, and the number of initial disagreements was not recorded. The final ratings therefore reflect consensus judgements whose reproducibility by independent reviewers has not been quantified; this is acknowledged as a limitation in Section 4.3. The distribution of ratings across the corpus is reported in Section 3.

F. Data Synthesis

Synthesis was conducted as a narrative review; no meta-analysis was performed as heterogeneity in evaluation datasets, metric choices and experimental conditions precluded quantitative pooling. Evidence from the included studies was organised around the four research questions; findings were grouped under

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

nine synthesis themes (T1-T9) that emerged iteratively during extraction. Theme assignments were confirmed at structured synthesis checkpoints after every ten records and adjusted where extraction of subsequent records warranted reclassification.

Within each theme, studies were compared on the basis of dataset identity, evaluation protocol and reported metric values. Comparative performance claims were conditioned on shared evaluation conditions, figures from studies using different datasets or non-standard class distributions were not pooled into aggregate rankings. Where studies reported results on the same benchmark dataset under comparable protocols, performance was compared directly where conditions differed, findings were synthesised descriptively and the source of variation identified.

III. RESULTS AND DISCUSSION

A. Characteristics of Included Studies

Database searches retrieved 3,738 records across five sources where deduplication removed 1,009 records, the remaining 2,729 underwent title and abstract screening of which 2,669 were excluded. The 60 full-text articles retrieved for eligibility assessment yielded 50 included studies after 10 were excluded; exclusion reasons are detailed in Figure 1.

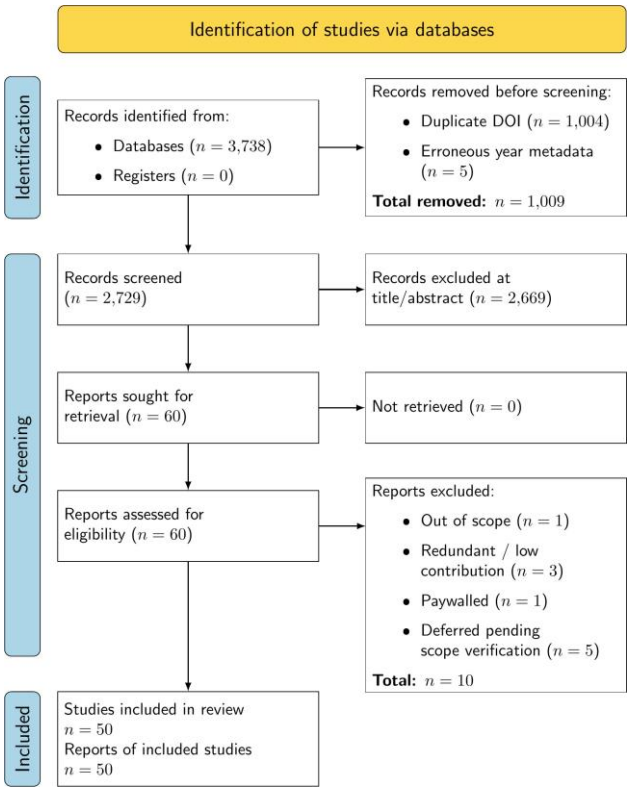


Figure 1. PRISMA 2020 Flow Diagram

The 50 included studies span 2017 to 2026. 42 (84%) appeared from 2022 onwards; the corpus is therefore weighted towards the most recent phase of ML-based IDS research. Quality assessment assigned five studies a Strong rating, 44 an Adequate rating, and one a Weak rating. The characteristics of all included studies are listed in Table 6 (Appendix A).

B. Machine Learning Technique Taxonomy (RQ1)

The 50 included studies were assigned across nine synthesis themes; individual papers appear in more than one theme where their primary method spanned multiple technique families. Table 3 summarises the

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

distribution and Figure 2 presents the hierarchical taxonomy of technique families and representative algorithms. Figure 2 organises the nine themes into three paradigm groups; the distribution within those groups is discussed below. Classical ML (T1, $n = 23$) and deep learning (T3, $n = 24$) together account for 64% of theme assignments and represent the two dominant paradigms in the corpus. Within T1, random forest is the most frequently applied classifier, appearing in 14 studies (Singh et al., 2021; Megantara and Ahmad, 2021; Le Jeune et al., 2021; Zhang et al., 2020; Zou et al., 2023; Raza et al., 2023; Abdulboriy and Shin, 2024; Ma et al., 2025; Singh and Ujjwal, 2023; Alqahtani et al., 2025; Kurniabudi et al., 2022; Korium et al., 2024; Zhao et al., 2023; Wang et al., 2025), This dominance holds across NSL-KDD, UNSW-NB15, CICIDS-2017, and Net-Flow benchmarks. Support vector machines constitute the second most common T1 method, present in nine studies (Le Jeune et al., 2021; Zou et al., 2023; Abdulboriy and Shin, 2024; Singh and Ujjwal, 2023; Wang, 2026; Rodríguez et al., 2022; Tawfik, 2024; Alqahtani et al., 2025; Korium et al., 2024), and decision tree variants appear in eleven (Singh et al., 2021; Megantara and Ahmad, 2021; Le Jeune et al., 2021; Ahmed et al., 2024; Abdulboriy and Shin, 2024; Ro-dríguez et al., 2022; Tawfik, 2024; Alqahtani et al., 2025; Yousuf and Mir, 2025; Kurniabudi et al., 2022; Wang et al., 2025).

Within T3, long short-term memory networks and their variants are the most frequently applied deep architecture, present in thirteen studies (Laghrissi et al., 2021; Elsaid et al., 2024; Wang et al., 2026; Raza et al., 2023; Sun et al., 2025; Yoshimura et al., 2022; Yao et al., 2023; Al-Daweri et al., 2025; Alshammari et al., 2026; Wang et al., 2026; Pinto et al., 2026; Hossain, 2025; Xu et al., 2025). Convolutional neural networks appear in twelve studies (Shahzad et al., 2022; Wang et al., 2026; Yao and Lin, 2025; Hossain, 2025; Raza et al., 2023; Sun et al., 2025; Yoshimura et al., 2022; Zhang et al., 2025; Xu et al., 2025; Alshammari et al., 2026; Zhu et al., 2025; Aljameel et al., 2025), and autoencoder architectures in five (Torabi et al., 2023; Zhang et al., 2020; Yao et al., 2023; Zhao et al., 2023; Al-Daweri et al., 2025). Transformer-based methods appear in two studies (Luay et al., 2026; Xu et al., 2025), both published in 2025-2026, in line with the broader adoption of attention mechanisms across deep learning tasks.

Table 3. Distribution of Included Studies across Synthesis Themes

Theme	Technique Family	Assignments (*n*)
T1	Classical ML (SVM, RF, k-NN, NB, DT, ELM)	23
T2	Ensemble and boosting (XGBoost, LightGBM, stacking, voting)	3
T3	Deep learning (CNN, LSTM, AE, GAN, Transformer)	24
T4	Hybrid and multi-stage architectures	5
T5	Benchmark, feature engineering, and data augmentation	7
T6	Deployment, edge, and real-time systems	7
T7	Quantum and novel computational paradigms	1
T8	ICS/CPS-specific intrusion detection	1
T9	Graph-based detection (GNN, knowledge graphs)	3
Total theme assignments (papers may appear in multiple themes)		74

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

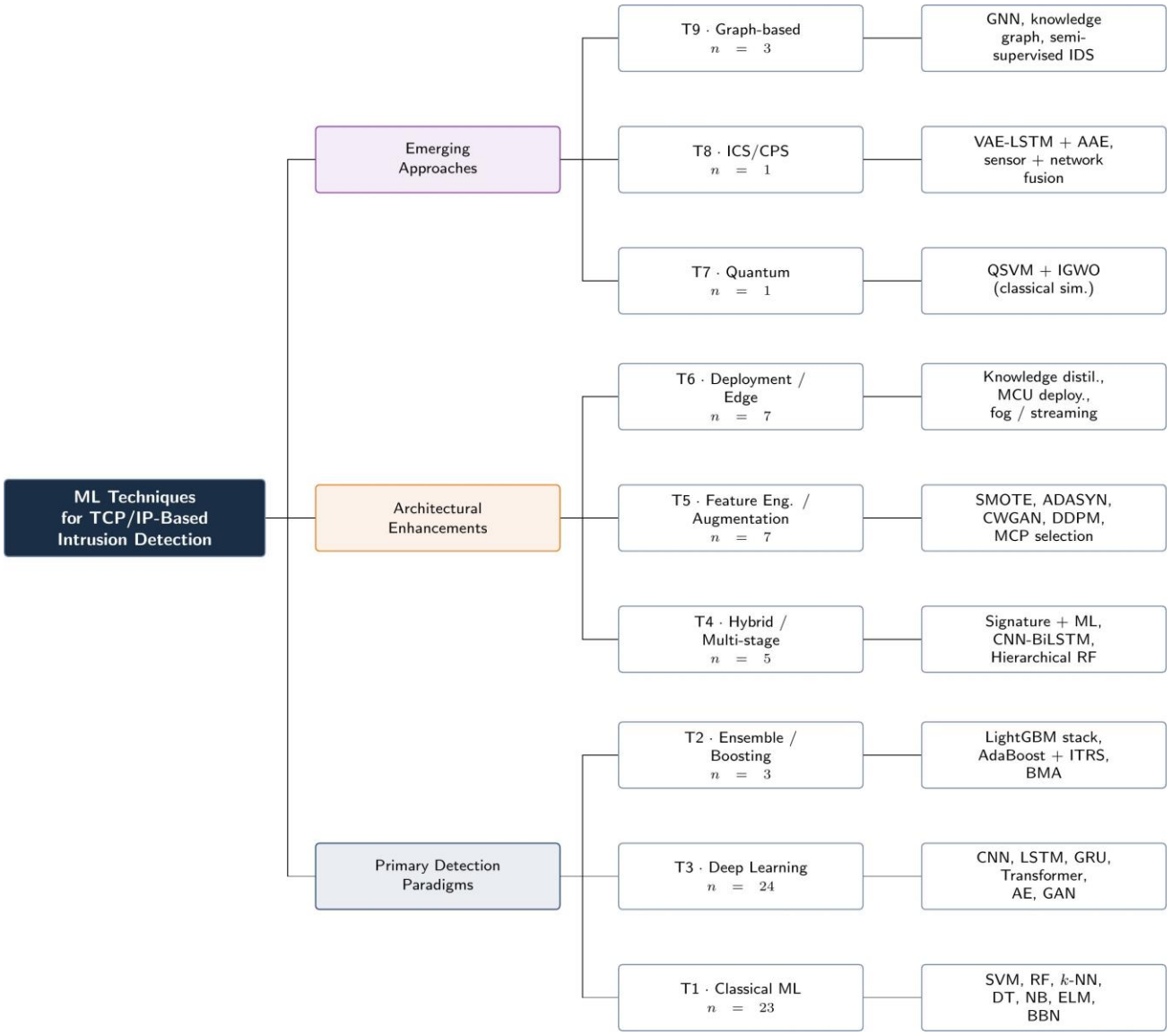


Figure 2. Hierarchical Taxonomy of Machine Learning Technique Families Applied in the Study

Themes are organized into three paradigm groups: Primary Detection Paradigms (T1, T3, T2), Architectural Enhancements (T4, T5, T6), and Emerging Approaches (T7, T8, T9), but analytically important cluster. Ahmed et al. (2024) combine stacking, Bayesian model averaging, and a conditional ensemble; Maksimovic´ et al. (2024) apply AdaBoost with a custom over-voting scheme; Yousuf and Mir (2025) stack Light GBM with a logistic regression meta learner. All three papers report marginal gains over their best individual base classifier, this pattern (strong individual learners leaving limited headroom for ensemble improvement on these bench marks) recurs across all three T2 papers and constitutes a consistent finding rather than an isolated anomaly. Hybrid and multi-stage architectures (T4, $n = 5$) combine two or more distinct detection paradigms in sequence. Singh et al. (2021) integrate a signature module with a meta-ensemble anomaly detector; Shahzad et al. (2022) chain a 1D CNN feature extractor with an LSTM classifier; Zou et al. (2023) precede an ELM with hierarchical clustering; Raza et al. (2023) apply a two-stage CNN BiLSTM pipeline; Wang et al. (2025) use a hierarchical RF with multiresolution feature importance. Each T4 paper targets a coverage gap that single paradigm components cannot address alone. Feature engineering, benchmark studies and data augmentation (T5, $n = 7$) cut across the corpus as a secondary methodological layer. Four studies apply generative augmentation to address minority class scarcity (Zhang et al., 2020, 2025; Zhu et al., 2025; Aljameel et

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

al., 2025), two examine feature selection pipelines as the primary contribution (Ma et al., 2025; Al-Daweri et al., 2025), one surveys feature selection methods across the IIoT IDS literature (Francis et al., 2026).

Deployment oriented studies (T6, $n = 7$) report detection in resource constrained, streaming or edge environments. Ahmed et al. (2020) and Abdulboriy and Shin (2024) address online and streaming detection; Zhao et al. (2023) demonstrate MCU deployment for IDS; Yao and Lin (2025) apply knowledge distillation for model compression; Yoshimura et al. (2022), Tawfik (2024) and Yao et al. (2023) evaluate edge or fog node scenarios. Three thin themes emerged from extraction with one or two papers each. Elsedimy et al. (2024) constitutes the sole T7 entry. The paper applies a quantum support vector machine simulated on classical hardware, no quantum speedup is demonstrated and the result is best interpreted as a Proof of Concept (PoC). Pinto et al. (2026) alone occupies T8, the paper fuses physical sensor data from a water treatment plant with TCP/IP Modbus traffic in a multimodal CPS anomaly detector. T9 covers graph based detection across three studies (Eltanbouly et al., 2025; Mananayaka and Chung, 2026; Zheng et al., 2025); Eltanbouly et al. (2025) surveys the area while Mananayaka and Chung (2026) and Zheng et al. (2025) contribute primary graph based IDS implementations using knowledge graphs and GNNs respectively. The small paper counts in T7, T8, and T9 indicate genuine under investigation rather than a search gap and these subfields have not yet produced a critical mass of comparable primary evaluations.

C. Benchmark Datasets (RQ2)

The 50 included studies drew on 19 distinct datasets. Three benchmarks dominated the evaluation space: NSL-KDD appeared in 19 studies, CICIDS-2017 in 18 and UNSW-NB15 in 15. The remaining 16 datasets each appeared in three or fewer studies and eight appeared in only one. Table 4 summarises the ten most frequently used benchmarks and Figure 3 shows the full frequency distribution. The concentration of evaluation on three legacy benchmarks has a direct consequence for the comparability of reported results. NSL-KDD, CICIDS-2017 and UNSW-NB15 were all generated under controlled laboratory or simulation conditions and do not capture the encrypted, high volume, heterogeneous traffic of contemporary TCP/IP networks (Eltanbouly et al., 2025). Studies evaluated on NSL-KDD are not directly comparable with those evaluated on CICIDS-2017 because attack taxonomies, feature sets and class distributions differ substantially across the two benchmarks (Luay et al., 2026). Across the 50 papers, only seven used more than one benchmark in a way that permits cross dataset performance comparison (Le Jeune et al., 2021; Ma et al., 2025; Luay et al., 2026; Zhao et al., 2023; Hossain, 2025; Alshammari et al., 2026; Xu et al., 2025).

The CICIDS-2017 Wednesday subset is a particular source of distortion. Ten of the 18 studies citing CICIDS-2017 used the Wednesday subset in isolation; this restricted evaluation to five attack classes. The Heartbleed class contains only 11 labelled samples. F1 estimates carry high variance across random splits and per class comparisons between these studies are unreliable (Raza et al., 2023; Sun et al., 2025; Alqahtani et al., 2025). A further structural problem is the mismatch between study framing and benchmark origin. Francis et al. (2026) found that 63.9% of IIoT-IDS papers in their reviewed sample used benchmarks not collected from IIoT environments. The same pattern appears in this corpus: studies framed as IIoT or ICS/CPS evaluations predominantly used CICIDS-2017 or NSL-KDD, neither of which was collected from IoT devices or industrial control systems (Elsedimy et al., 2024; Singh and Ujjwal, 2023; Kurniabudi et al., 2022; Tawfik, 2024). The NF-v2 NetFlow family constitutes the only standardised multi dataset framework in the corpus enabling controlled cross study comparison. The family imposes a common 43 feature schema across five sub datasets; performance differences can therefore be attributed to dataset characteristics rather than feature engineering choices (Ma et al., 2025; Luay et al., 2026). Zhao et al. (2023) extend this framework to edge deployment on the NF-UNSW-NB15-v2 sub-dataset under MCU resource constraints. Three studies is an insufficient base from which to draw general conclusions about cross dataset generalisation, but the NF-v2 family provides the methodological template the field requires.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

Implications for researchers who continue to use these benchmarks. The evidence above does not argue that NSL-KDD, CICIDS-2017 and UNSW-NB15 should be abandoned. They are the common ground on which new methods can be compared with the 43 of 50 corpus studies that use them, and their public availability supports replication. It does mean that results on these datasets should be read, and reported, as benchmark-conditional evidence of detection capability and not as evidence of deployment readiness. A model that performs well on one legacy benchmark has shown that it can separate that dataset's laboratory or simulated traffic, which differs from contemporary encrypted, high-volume and heterogeneous TCP/IP traffic (Eltanbouly et al., 2025).

First, performance should be reported with metrics that survive class imbalance. Raw accuracy conceals minority class failure (Le Jeune et al., 2021), so macro-F1 and per-class precision and recall should be the primary results and accuracy secondary. Second, per-class results for rare classes should carry their sample counts and a measure of variance across random splits or seeds. The Heartbleed class in CICIDS-2017 contains only 11 labelled samples, and F1 estimates on so few cases are too unstable to rank methods. Third, the data subset and split protocol should be stated explicitly. Ten of the 18 CICIDS-2017 studies used the Wednesday subset alone, which covers five attack classes, so their results are not comparable with full-dataset evaluations and should not be set against them. Fourth, claims should be scoped to what the benchmark can support. NSL-KDD and CICIDS-2017 were not collected from IoT or industrial control environments, so studies framed around those settings should either evaluate on data collected there or state that the framing is motivational and not demonstrated (Francis et al., 2026). KDD Cup 1999, in which about 78% of records are duplicates, should be deduplicated before use, or NSL-KDD used instead. Fifth, the label provenance of CICIDS-2017 should be recorded, given the label errors documented by Eltanbouly et al. (2025), and any correction applied should be reported. Sixth, evaluation on a single benchmark should be treated as a minimum and not as a conclusion. Cross-dataset claims require datasets that share a feature schema, which the NF-v2 family provides; results obtained on datasets with different feature extraction should be presented as separate within-dataset results and not as evidence of generalisation.

A further implication concerns minority classes. Because U2R and R2L attacks lack distinctive flow-level temporal signatures (Luay et al., 2026), improving their detection on NSL-KDD or UNSW-NB15 through oversampling or generative augmentation alone should not be interpreted as a capability that will carry over to deployment. Studies that use augmentation should report whether the discriminative information for the class is present in the available features, for example by testing richer features such as interarrival timing, and should distinguish gains from sample balancing from gains from new feature information.

Finally, results from these benchmarks should be compared only under matched conditions: the same dataset, the same split protocol and the same imbalance handling. Published figures obtained under different conditions should not be pooled into rankings, as this review does not pool them in Section 3.4. Researchers who follow these practices can continue to use the legacy benchmarks without overstating what the results show, and their findings will remain comparable with the existing literature.

D. Comparative Performance (RQ3)

Direct performance comparisons across studies are valid only when evaluation conditions are held constant. The three conditions that most frequently differ across the corpus are the benchmark dataset, the train test split ratio and whether class imbalance was corrected before training. All performance figures cited in this subsection are conditioned on shared evaluation settings, cross dataset comparisons from published tables are identified as such and excluded from quantitative claims. Where ranges are reported (such as F1 0.00-0.65 for U2R), they reflect between study heterogeneity under varying evaluation conditions, not statistical uncertainty around a pooled estimate, they cannot be interpreted as confidence intervals.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

Table 4. Most Frequently used Evaluation Datasets ($n \geq 2$).

Dataset	*n*	Known limitations
NSL-KDD	19	Derived from KDD Cup 1999; traffic synthesised from a 1998 DARPA simulation run by the Defense Advanced Research Projects Agency (DARPA); does not capture modern TCP/IP protocol behaviour or contemporary attack vectors.
CICIDS-2017	18	Label errors documented by Eltanbouly et al. (2025); Wednesday subset used in isolation by ten studies, restricting attack type coverage to five classes; Heartbleed minority (11 samples) produces high variance per class F1.
UNSW-NB15	15	Synthetic traffic generated via IXIA Perfect Storm; U2R and R2L classes are severely underrepresented, producing near-zero recall in most studies.
KDD Cup 1999	3	Approximately 78% duplicate records; retained only in Teng et al. (2018), Wang (2026), and Aljameel et al. (2025), each alongside more recent benchmarks.
NF-v2 family	3	Standardised 43-feature NetFlow schema across five sub-datasets; the only benchmark family enabling consistent cross-dataset comparison in this corpus (Ma et al., 2025; Luay et al., 2026; Zhao et al., 2023).
BoT-IoT	3	IoT botnet traffic; 53 million records; subset sampling undescribed in two of three studies; class imbalance unaddressed in Elsedimy et al. (2024).
CICIDS-2018	3	Extended CICIDS-2017 with additional attack types; used by Xu et al. (2025), Alshammari et al. (2026), and Zhu et al. (2025).
CICIoT2023	2	2.6 million samples; 34 attack classes; largest by class count in the corpus; used by Hossain (2025) and Alshammari et al. (2026).
CIC IoT-DIAD 2024	2	414,990 samples; 33 attacks; 105 IoT device types; used by Hossain (2025) and Alshammari et al. (2026).
TON-IoT	2	Heterogeneous IoT and IIoT traffic; used within the NF-v2 family by Ma et al. (2025) and Luay et al. (2026).

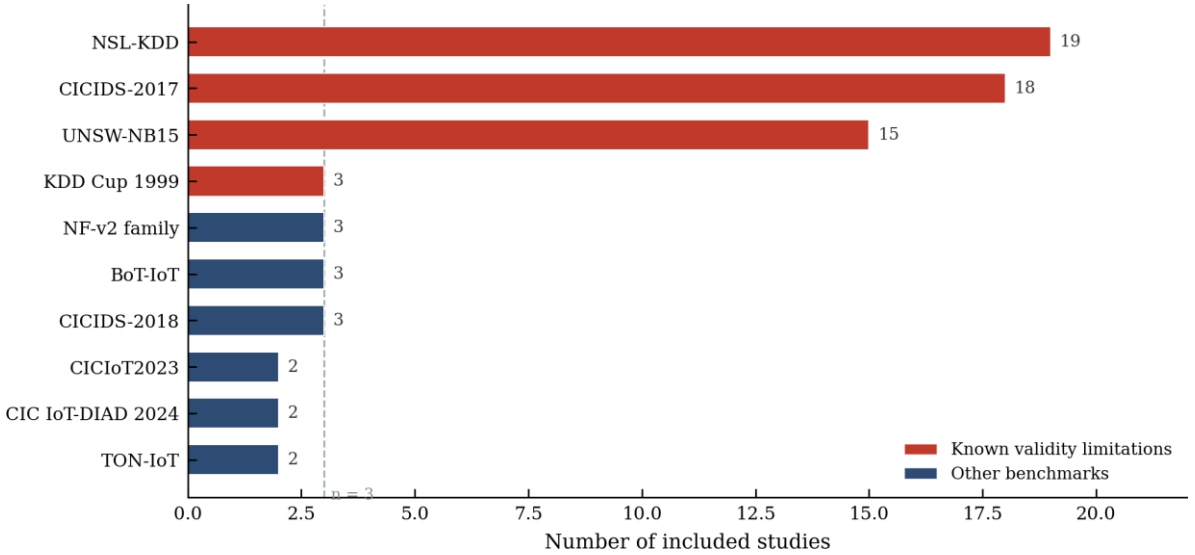


Figure 3. Frequency of Evaluation Datasets across the 50 included Studies.
*Datasets with known validity limitations are shown in red, all others in navy. The dashed line marks $n = 3$.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

(a). Accuracy Inflation and Metric Reliability

Raw accuracy is an unreliable primary metric for TCP/IP intrusion detection benchmarks. Le Jeune et al. (2021) demonstrate this formally: under their Detection Score (DS) / Identification Score (IS) framework, a model achieving DS 0.91 on CICIDS-2017 may simultaneously achieve IS 0.93 or IS 0.61 depending on whether it detects minority attack classes. An SVM achieving overall accuracy above 95% on CICIDS-2017 produces IS 0.61 on the same data, near complete failure on rare attack types follows (Le Jeune et al., 2021). Eltanbouly et al. (2025) independently confirm that label errors in CICIDS-2017 inflate accuracy for models that memorise majority class boundaries. F1-macro is the metric least susceptible to majority-class inflation and is used as the primary comparative metric throughout this subsection.

(b). Classical ML Performance

Random forest is the strongest classical ML classifier across the benchmarks represented in this corpus. On CICIDS-2017, Sun et al. (2025) report RF F1-macro-0.931; on UNSW-NB15, Sun et al. (2025) report RF F1-macro-0.869; on NSL-KDD, Zhang et al. (2020) report RF at 87.94% accuracy against a hard test subset. The most controlled classical ML evaluation in the corpus is provided by Le Jeune et al. (2021), who report RF IS 0.93 on CICIDS-2017 under a unified feature-engineering protocol that holds pre-processing constant across all compared classifiers. SVM is the second most consistent performer on majority classes but the most sensitive to minority class scarcity; IS 0.61 in Le Jeune et al. (2021) against RF IS 0.93 on the same data represents the largest intra-benchmark gap between any two classifiers in the corpus.

(c). The Ensemble Paradox

Ensemble methods do not consistently outperform their strongest base classifier on these benchmarks. Ahmed et al. (2024) report that LightGBM alone achieves 99.7% accuracy and F1-macro 0.986 on CICIDS-2017, while their stacking ensemble reaches 99.8% accuracy and F1-macro-0.989, a gain of 0.003 F1-macro at five times the training cost. Yousuf and Mir (2025) observe the same pattern: LightGBM alone at F1-macro 0.986 and stacking at 0.989; the marginal gain is not statistically tested by the authors. Maksimovic et al. (2024) report that AdaBoost with their ITRS over voting scheme improves binary DDoS accuracy by less than 0.15% over AdaBoost alone. The ensemble paradox arises because the dominant benchmarks are sufficiently separable for strong individual learners; ensemble diversity gains are masked by ceiling effects on these datasets.

(d). Deep Learning Performance

CNN-BiLSTM architectures with attention mechanisms represent the current performance ceiling on the principal benchmarks. Sun et al. (2025) report their CNN-BiLSTM with cross-modal attention achieving F1-macro 0.974 on full CICIDS-2017 (15 classes) and F1-macro 0.941 on UNSW-NB15 (10 classes). Adding cross-modal attention over naive CNN+LSTM concatenation contributes +0.016 F1-macro on CICIDS-2017 and +0.040 on UNSW-NB15 (Sun et al., 2025). Alshammari et al. (2026) report F1-macro 0.981 on CIC-IoT2023 (34 classes); ablation attributes +0.011 to self-attention and a further +0.007 to cross-attention. Knowledge distillation reduces model parameters while preserving accuracy: Yao and Lin (2025) report their compressed model achieving accuracy 0.93 and AUC 0.98 on CICIDS-2017 with only 19,000 parameters, which is both smaller and more accurate than the RF (113,000 parameters, AUC 0.91) and SVM (42,000 parameters, AUC 0.88) baselines on the same data.

(e). The Temporal Feature Advantage

Interarrival time features provide the largest single performance gain documented in the corpus for multiclass detection. Luay et al. (2026) compare the NF3 temporally enriched feature set against the standard NF-v2 schema across five datasets and three sequential architectures. The multiclass macro F1 improvement is +27.3% for the Transformer, +24.1% for LSTM, and +19.8% for 1D-CNN averaged across datasets. The mechanism is specific: DoS Hulk, Recon-naissance, and Botnet C2 attacks each

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

exceed 30% macro F1 improvement under NF3, because these attack classes carry interarrival timing signatures that static feature vectors discard (Luay et al., 2026). Application layer attacks (web attacks, SQL injection) show under 3% NF3 gain; temporal features address network-flow-detectable threats rather than payload-dependent ones. Binary classification gains from temporal features are smaller: +2.1% F1 for the Trans-former, +1.8% for LSTM. The asymmetry between binary and multiclass gains is the strongest evidence in the corpus that aggregate accuracy conceals class level capability differences. Figure 4 quantifies this asymmetry.

(f). Minority Class Failure

Minority class failure is the most consistent and cross cutting finding in the corpus. The pattern appears across nine studies and six datasets, and is independent of classifier family. Table 5 documents the per class F1 scores for the four most problematic attack categories across studies that report them. Figure 5 plots the F1 range for each class across all studies that report per class results. The U2R class in NSL-KDD is the most severe case. Without augmentation, U2R F1 is effectively zero in most studies; CWGAN augmentation raises U2R accuracy to 42.50% against an RF baseline of 15.82% and an SVM baseline of 8.35% (Zhang et al., 2020). Al-Daweri et al. (2025) apply ADASYN oversampling to NSL-KDD and report R2L F1 72.4% and U2R F1 48.9%, improved over no oversampling baselines but still the two lowest classes after oversampling. The mechanistic explanation is provided by Luay et al. (2026): U2R and R2L attacks operate at the application layer and do not produce distinctive network flow temporal signatures; no amount of oversampling addresses the fundamental absence of discriminative TCP/IP level features for these classes. Generative augmentation addresses sample scarcity but not feature poverty; this distinction is a finding with direct implications for benchmark design.

(g). Cross-Domain Generalisation

Cross domain performance, defined here as training on one dataset and testing on another without fine-tuning, drops substantially relative to within domain results across all studies that report it. Mean F1-macro 0.891 across 12 cross dataset transfer permutations is reported by Zhao et al. (2023) for their DI-NIDS framework, against F1-macro 0.743 for a standard CNN under the same conditions; the gap of 0.148 F1-macro is attributable to their domain-adversarial normalisation layer. Luay et al. (2026) report mean detection rate 0.9832 and false alarm rate (FAR) 0.0035 across 20 cross-dataset pairs within the NF3 family, but this requires the shared feature schema that NF-v2/NF3 provides. No study in the corpus demonstrates reliable cross domain generalisation between datasets that do not share a common feature extraction protocol; benchmark specific feature engineering is therefore the primary barrier to deployment portability.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

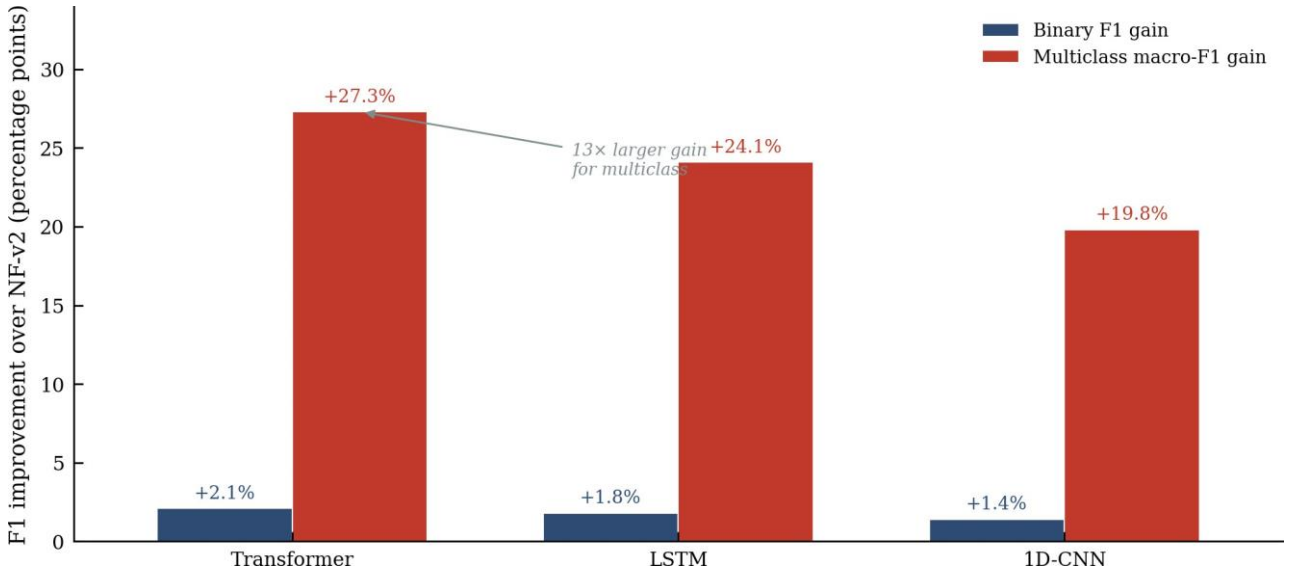


Figure 4. F1 improvement from NF3 Temporal Features over the Standard NF-v2 Schema
*Binary gains are minor (+1.4 - 2.1 percentage points); multiclass macro-F1 gains are 13 × larger (+19.8 - 27.3 percentage points).

Table 5. Per-class F1 for Persistent Minority Attack Categories.

Attack class	Dataset	F1 range	Studies
U2R	NSL-KDD	0.00-0.65	(Altamimi and Abu Al-Haija, 2024); (Teng et al., 2018); (Zhang et al., 2020); (Zou et al., 2023); (Zhu et al., 2025)
R2L	NSL-KDD	0.35-0.72	(Teng et al., 2018); (Zhang et al., 2020); (Zou et al., 2023); (Abdulboriy and Shin, 2024); (Al-Daweri et al., 2025)
Heartbleed	CICIDS-2017	0.76-0.91	(Raza et al., 2023); (Sun et al., 2025); (Alqahtani et al., 2025); (Yousuf and Mir, 2025); (Al-Daweri et al., 2025)
Infiltration	CICIDS-2017	0.00-0.77	(Sun et al., 2025); (Al-Daweri et al., 2025)

*Scores below 0.75 are diagnostic of Minority Class Failure

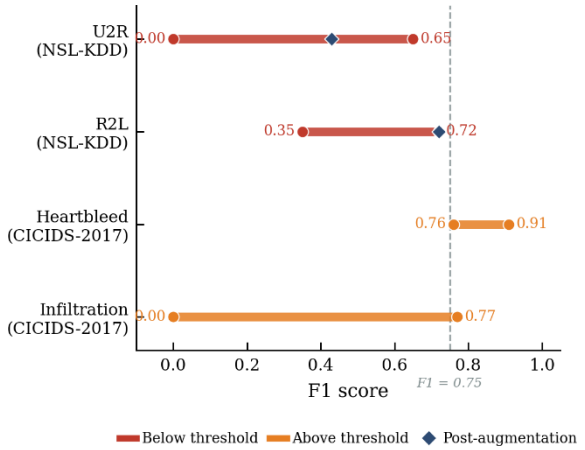


Figure 5. Per-class F1 Score Ranges for Persistent Minority Attack Categories.

Bars show the minimum -maximum range across reporting studies; diamond markers show the best post

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

augmentation result where available. The dashed line at F1 = 0.75 marks the failure threshold.

(h). Deployment Efficiency

Inference latency and model size are reported by fewer than one third of the included studies, a gap that Eltanbouly et al. (2025) and Francis et al. (2026) independently identify as systemic. Among studies that do report latency, figures range from 0.3ms per sample for a lightweight CNN on CICIDS-2017 (Xu et al., 2025) to 891ms per micro batch at one million flows per second for a distributed Spark RF (Ahmed et al., 2020). Zhao et al. (2023) demonstrate MCU deployment with 18K parameters, 12 KB RAM usage, and a model size of 72 KB, the most constrained resource profile in the corpus. Yao and Lin (2025) achieve inference time 0.07 seconds at 5,000 samples through knowledge distillation, against 22.8 seconds training time. The absence of a standard latency reporting protocol across the corpus makes these figures non-comparable, the studies that report latency use different hardware, different sample sizes and different measurement points in the pipeline.

E. Open Challenges (RQ4)

Eight principal gaps emerge from the corpus, each evidenced by convergent findings across multiple independent studies.

(a). Minority Class Detection

The failure of ML-based IDS to detect rare attack classes is the most pervasive and best evidenced gap in the corpus. U2R recall on NSL-KDD is effectively zero without augmentation across five studies (Altamimi and Abu Al-Haija, 2024; Teng et al., 2018; Zhang et al., 2020; Zou et al., 2023; Zhu et al., 2025); ADASYN oversampling raises U2R F1 to 48.9%, still the lowest class after correction (Al-Daweri et al., 2025). Luay et al. (2026) identify the mechanistic cause: U2R and R2L attacks carry application layer signatures that are absent from TCP/IP flow level feature vectors. Generative augmentation raises the detection floor (Zhang et al., 2020) but does not address the underlying feature poverty. No study in the corpus proposes a feature extraction protocol that captures application layer semantics within a network flow framework, this constitutes an open design gap.

(b). Benchmark Standardisation

The absence of a shared evaluation standard makes cross study performance comparison structurally impossible for the majority of papers in the corpus. Nineteen distinct datasets appear across 50 studies. Three legacy benchmarks (NSL-KDD, CICIDS-2017, UNSW-NB15) carry documented validity problems: outdated traffic profiles, label errors and severe class imbalance (Eltanbouly et al., 2025). Only the NF-v2/NF3 family imposes a common 43 feature schema enabling controlled cross study comparison and only three corpus papers use it (Ma et al., 2025; Luay et al., 2026; Zhao et al., 2023). Francis et al. (2026) find that 63.9% of IIoT IDS papers use benchmarks not collected from IIoT environments, the same mismatch appears in this corpus for IoT and ICS framed studies.

(c). Real-Time Latency Reporting

Inference latency is reported by fewer than one third of included studies. Eltanbouly et al. (2025) find only 31% of surveyed NIDS papers report inference time; Francis et al. (2026) report 19.4% in the IIoT IDS literature. Among the 14 corpus papers that do report latency, figures span three orders of magnitude: from 0.07 seconds at 5,000 samples under knowledge distillation (Yao and Lin, 2025) to 891 milliseconds per micro-batch at one million flows per second under distributed Spark processing (Ahmed et al., 2020). No standard latency reporting protocol exists across the field, and no study specifies a target latency against a deployment SLA. A model that achieves F1-macro 0.98 at 2,000 milliseconds per sample cannot be deployed in real-time network monitoring, yet this constraint is absent from 70% of the papers that claim practical applicability.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

(d). Concept Drift Adaptation

Only one primary study in the corpus directly addresses concept drift in operational IDS: Abdulboriy and Shin (2024), who apply adaptive random forest with ADWIN drift detection on UNSW-NB15 and NSL-KDD streams; the approach achieves 350× model update speedup over batch retraining. Concept drift is identified as an open gap by Eltanbouly et al. (2025), Francis et al. (2026), and a further eight individual primary studies (Wang, 2026; Tawfik, 2024; Yao et al., 2023; Alshammari et al., 2026; Al-Daweri et al., 2025; Kurniabudi et al., 2022; Korium et al., 2024; Sun et al., 2025), none of which address it in their own experimental design. Static models trained on fixed benchmark snapshots cannot detect novel attack variants that emerge after training; the absence of drift aware evaluation in 49 of 50 studies is the clearest gap between the corpus and production deployment requirements.

(e). Adversarial Robustness

No study in the corpus evaluates IDS performance against adversarial inputs designed to evade detection. Adversarial robustness is flagged as an open gap by Eltanbouly et al. (2025), Hossain (2025), and six primary studies (Tawfik, 2024; Yao et al., 2023; Alshammari et al., 2026; Singh and Ujjwal, 2023; Korium et al., 2024; Wang, 2026), all of which identify it as a known weakness without evaluating it. All performance figures reported in the corpus are measured on clean, unperturbed test sets; no study reports detection rate under feature space adversarial perturbation. Deep learning models are known to be sensitive to gradient-based attacks in other domains, and no corpus paper tests for this vulnerability in an IDS setting.

(f). Cross-Domain Generalisation

Generalisation from one network environment to another without retraining or fine-tuning is demonstrated by only two corpus papers. Zhao et al. (2023) show that domain adversarial normalisation raises cross dataset F1-macro from 0.743 to 0.891 in 12 transfer permutations within the NF-v2 family. Luay et al. (2026) demonstrate mean detection rate 0.9832 across 20 cross dataset pairs within the NF3 feature schema, but this requires the shared feature extraction that NF-v2 provides. No study demonstrates generalisation between datasets with different feature extraction protocols; every model in the corpus is therefore implicitly dataset specific. Production IDS deployments must transfer across hardware vendors, network topologies, and traffic profiles; the current evaluation paradigm does not test for this.

(g). Explainability

Post hoc explainability methods (SHAP and LIME) appear in two studies (Ahmed et al., 2024; Zhang et al., 2025) as secondary contributions rather than primary evaluation targets. Keshavamurthy and Kulkarni (2025) provide the only example of a model with intrinsic transparency by design: Fuzzy Cognitive Maps produce interpretable causal graphs without post hoc overhead. Pinto et al. (2026) explicitly acknowledge XAI for ICS anomaly detection as an open direction. Regulatory frameworks for critical infrastructure increasingly require explainable automated decisions, yet 47 of 50 corpus papers provide no mechanism for a human analyst to understand or audit individual detection outcomes.

(h). Edge and Embedded Deployment

The gap between claimed and demonstrated edge deplorability is wide. Seventeen studies claim IoT, IIoT, edge, or fog deployment as motivation, but only four validate on constrained hardware: Zhao et al. (2023) on an MCU (12 KB RAM, 72 KB model); Yao and Lin (2025) on a knowledge-distilled model with 19,000 parameters; Torabi et al. (2023) on a cloud-edge simulation; and Ahmed et al. (2020) on a distributed Spark cluster. The remaining thirteen IoT or edge-framed studies evaluate exclusively on server class hardware; GPU requirements range from an RTX 3080 Ti to an A100. Francis et al. (2026) find that only 4 of 36 reviewed IIoT IDS papers address real time latency on constrained hardware. The hardware

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

validation gap is the single most consistent discrepancy between paper framing and evaluation protocol in the corpus.

(i). Principal Findings

This systematic review synthesised evidence from 50 primary studies on machine learning techniques for TCP/IP-based intrusion detection, addressing four research questions across nine synthesis themes. The answer to RQ1 is that the field is split between two dominant paradigms of near equal representation: classical ML, led by random forest across 23 studies, and deep learning, led by LSTM and CNN variants across 24 studies. Neither paradigm has established sustained superiority over the other, the observed performance differences between them are better explained by benchmark choice and evaluation protocol than by architectural advantage. Ensemble and boosting methods (T2, $n = 3$) consistently fail to outperform their strongest base classifier by more than 0.003 F1 macro, a finding that holds across three independent studies using different ensemble strategies. Thin themes T7, T8, and T9 (quantum computation, ICS/CPS detection, and graph-based IDS) each contain one to three studies and represent areas of genuine under-investigation rather than search gaps. The answer to RQ2 is that benchmark adequacy is the field's most critical unresolved problem. Three legacy datasets (NSL-KDD, CICIDS-2017, UNSW-NB15) account for the evaluation of 43 of 50 included studies, yet each carries documented validity limitations: NSL-KDD derives from a 1998 DARPA simulation; CICIDS-2017 contains known label errors (Eltanbouly et al., 2025); and UNSW-NB15 severely underrepresents its U2R and R2L classes. Only the NF-v2/NF3 NetFlow family provides a standardised multi-dataset schema enabling controlled cross study comparison, and only three corpus papers use it. The mismatch between study framing and benchmark origin is quantified by Francis et al. (2026), who find that 63.9% of IIoT IDS papers use benchmarks not collected from IIoT environments; the same pattern holds in this corpus.

Researchers who continue to use these benchmarks should therefore report macro-F1 and per-class results with variance, state the subset and split protocol used, scope deployment claims to the origin of the data, and use the NF-v2 family when cross-dataset claims are made (Section 3.3). The answer to RQ3, evidenced in full in Section 3.4, is that raw accuracy is an insufficient comparative metric for TCP/IP IDS: it conceals minority class failure and inflates apparent performance on class imbalanced benchmarks. The more consequential implication of RQ3 is that temporal interarrival features are the single largest untapped performance lever in the corpus: NF3 enrichment raises multiclass macro F1 by up to 27.3 percentage points (Luay et al., 2026) while binary accuracy changes by under 2.2 points on the same data. This asymmetry means that published accuracy figures systematically understate the benefit of richer feature representations for the attack classes that matter most. The minority class failure pattern is structural, not incidental: its root cause is featuring poverty at the TCP/IP flow level, not sample scarcity, and no augmentation strategy addresses that. The answer to RQ4 is eight convergent gaps, none of which is addressed by more than two studies in the corpus: minority class feature representation, benchmark standardisation, inference latency reporting, concept drift adaptation, adversarial robustness, cross-domain generalisation, explainability, and hardware validated edge deployment. Of these, concept drift adaptation and adversarial robustness are the most acute: each is flagged as critical by multiple surveys and primary papers, yet addressed experimentally by at most one study in the corpus.

(j). Positioning Against Prior Reviews

This review converges with prior IDS surveys on two findings and diverges on two others. The dominance of CICIDS-2017 and NSL-KDD as evaluation benchmarks aligns with surveys published through 2024 (Eltanbouly et al., 2025); benchmark lock-in is a structural feature of the field rather than a sampling artefact. The prevalence of minority class failure for U2R, R2L, and Heartbleed categories has been documented in prior syntheses (Eltanbouly et al., 2025; Francis et al., 2026). Two divergences carry more analytical weight than the con-vergences. First, evidence from three independent ensemble studies in this

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

corpus supports the finding that ensemble methods do not reliably outperform strong individual classifiers on these benchmarks, a result that challenges the conventional expectation of ensemble diversity gains. The pattern is limited to three studies and should be tested on harder benchmarks before a general claim is warranted, but it aligns with ceiling-effect arguments in the broader ML ensemble literature. Second, to the best of the authors' knowledge, no prior systematic review of the TCP/IP IDS literature has explicitly distinguished sample scarcity from feature poverty as separate causes of minority class failure. Prior surveys identify the failure pattern (Eltanbouly et al., 2025); this review identifies the mechanistic cause and explains why augmentation alone cannot fix it.

(k). Limitations

Four limitations constrain the generalisability of this review. First, search coverage was bounded by database access. Three databases (ACM Digital Library, Scopus, and Web of Science) were unavailable to the research team; studies indexed exclusively in these sources are absent from the corpus. The five searched databases cover the principal IEEE, Springer, Elsevier, Taylor & Francis, and biomedical literature, but the exclusion of ACM and Scopus represents a coverage constraint that may have missed conference-only publications and studies in venues outside the searched platforms. No independent index was available against which recall could be estimated, so the direction and size of any resulting omission cannot be quantified. This affects the comprehensiveness of the corpus rather than the quality of the included evidence, which was appraised separately (Section 2.6). It also constrains how corpus-level statements should be read: the dataset frequencies, the concentration on three legacy benchmarks, and gaps reported as absent in all 50 studies (such as adversarial robustness) describe the retrieved corpus. They indicate that a practice is rare in the retrieved literature and should not be read as proof that no study elsewhere has addressed it. Second, 1,538 of 2,729 screened records (56%) contained no abstract in their database exports; these records were screened on title alone, which increases the probability of false exclusions at the title/abstract stage. Third, the three thin themes (T7, T8 and T9) each contain three or fewer primary studies, synthesis claims within these themes rest on limited evidence and should be treated as provisional until larger bodies of directly comparable primary evaluations exist. Fourth, the quality ratings were reconciled by consensus without a reported inter-reviewer agreement statistic, so their reproducibility is not quantified, and the criteria in Table 8 are a purpose-built scheme and not a validated appraisal instrument; the quality distribution should be read as a descriptive indicator of evidence strength and not as a precise measurement.

(l). Future Research Directions

Eight scoped research directions follow directly from the gaps identified in Section 3.5

(a). Future studies should investigate feature extraction protocols that capture application layer attack semantics within TCP/IP flow level representations to address the feature poverty underlying U2R and R2L class failure that augmentation alone cannot resolve.

(b). The NF-v2/NF3 NetFlow family should be adopted as a standard evaluation schema across IDS research to enable controlled cross study comparison; its extension to IIoT and ICS/CPS traffic profiles would address the environment mismatch documented in 63.9% of IIoT IDS papers.

(c). A standard latency reporting protocol specifying hard-ware configuration, sample throughput, and target SLA should be established for IDS evaluation; the current absence of such a standard makes published latency figures non-comparable across studies.

(d). Concept drift adaptation mechanisms should be evaluated under realistic temporal traffic sequences rather than static benchmark snapshots; the single corpus study addressing drift (Abdulboriy and Shin 2024) demonstrates 350× update speedup; adaptive IDS is therefore feasible but systematically under investigated.

(e). Future IDS evaluations should include evasion attack protocols to test adversarial robustness; no

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

study in the reviewed corpus measures detection rate under feature space adversarial perturbation.

(f). Cross-domain generalisation should be evaluated be-tween datasets that do not share a common feature ex-traction protocol, extending beyond the NF-v2 family to test whether domain adversarial methods such as those in Zhao et al. 2023 transfer to heterogeneous traffic environments.

(g). Intrinsically transparent IDS models should be developed and compared against post-hoc XAI additions on the same data; the single corpus example of design time transparency (Keshavamurthy and Kulkarni 2025) demonstrates competitive accuracy without post-hoc overhead.

(h). Edge deployment claims should be validated on con-strained hardware; the four corpus studies that do so report feasible parameter counts and RAM footprints; hardware-validated edge IDS is achievable and should become a standard evaluation requirement.

IV. CONCLUSION

This review synthesised 50 primary studies on machine learning for TCP/IP-based intrusion detection. The evidence shows that the field is technically capable of high aggregate detection performance but structurally constrained by legacy benchmarks, unreported latency, and the persistent failure of all technique families on minority attack classes. The finding that temporal inter-arrival features raise multi-class macro-F1 by up to 27.3 percentage points over static flow features identifies a concrete and under exploited performance lever. The finding that sample augmentation cannot address minority class failure when feature poverty is the root cause reframes a decade long debate about oversampling in IDS research. The eight convergent gaps documented here are not incidental limitations of individual studies; they are structural properties of the field's evaluation paradigm, and addressing them is the necessary condition for producing IDS research that is reproducible, comparable, and deployable.

REFERENCES

Abdulboriy, A., & Shin, J. S. (2024). An incremental majority voting approach for intrusion detection system based on machine learning. *IEEE Access, 12*, 20058-20074. https://doi.org/10.1109/ACCESS.2024.3361041

Ahmed, A., Hameed, S., Rafi, M., & Mirza, Q. K. A. (2020). An intelligent and time-efficient DDoS identification framework for real-time enterprise networks: SAD-F: Spark based anomaly detection framework. *IEEE Access, 8*, 215160-215178. https://doi.org/10.1109/ACCESS.2020.3042905

Ahmed, U., Jiangbin, Z., Almogren, A., Khan, S., Sadiq, M. T., Altameem, A., & Rehman, A. U. (2024). Explainable AI-based innovative hybrid ensemble model for intrusion detection. *Journal of Cloud Computing: Advances, Systems and Applications, 13*, Article 196. https://doi.org/10.1186/s13677-024-00712-x

Al-Daweri, M. S., Abdullah, S., Khair, M., & Selamat, A. (2025). Intrusion detection using a bidirectional LSTM and attention mechanism with oversampling technique. *Scientific Reports, 15*, Article 6419. https://doi.org/10.1038/s41598-025-15631-3

Aldweesh, A., Derhab, A., & Emam, A. Z. (2020). Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. *Knowledge-Based Systems, 189*, Article 105124. https://doi.org/10.1016/j.knosys.2019.105124

Aljameel, S. S., Alabbad, D. A., Alzahrani, S. A., Alharbi, A. M., & Alzahrani, N. (2025). Diff-IDS: Enhancing intrusion detection using diffusion model-based oversampling for imbalanced datasets. *Computers, Materials & Continua, 82*(3), 3853-3874. https://doi.org/10.32604/cmc.2025.060357

Alqahtani, H., Sarker, I. H., Kalim, A., Hossain, S. M. M., Ikhlaq, S., & Hossain, S. (2025). Cyber intrusion detection using machine learning classification techniques. *Scientific Reports, 15*, Article 11516. https://doi.org/10.1038/s41598-025-32697-1

Alshammari, A., Aldrees, A., Alshamrani, M., & Al-Maleh, M. K. (2026). Enhancing IoT intrusion

detection through advanced deep learning and attention mechanisms. *Frontiers in Artificial Intelligence, 9*, Article 1825655. https://doi.org/10.3389/frai.2026.1825655

Altamimi, S., & Abu Al-Haija, Q. (2024). Maximizing intrusion detection efficiency for IoT networks using extreme learning machine. *Discover Internet of Things, 4*, Article 5. https://doi.org/10.1007/s43926-024-00060-x

Elsaid, S. A., Shehab, E., Mattar, A. M., Azar, A. T., & Hameed, I. A. (2024). Hybrid intrusion detection models based on GWO optimized deep learning. *Discover Applied Sciences, 6*, Article 531. https://doi.org/10.1007/s42452-024-06209-1

Elsedimy, E. I., Elhadidy, H., & Abohashish, S. M. M. (2024). A novel intrusion detection system based on a hybrid quantum support vector machine and improved Grey Wolf optimizer. *Cluster Computing, 27*(7), 9917-9935. https://doi.org/10.1007/s10586-024-04458-8 Advance online publication.

Eltanbouly, S., Zakraoui, J., Al-Ali, A., Belhi, A., Rahme, S., & Bouras, A. (2025). Network packet transformation approaches for intrusion detection systems: A survey. *IEEE Access, 13*, 107293-107312. https://doi.org/10.1109/ACCESS.2025.3581354 Advance online publication.

Francis, G. T., Souri, A., & Inanç, N. (2026). A systematic review of metaheuristic based feature selection strategies for cyber-attack detection in the IIoT. *Artificial Intelligence Review, 59*(2), Article 73. https://doi.org/10.1007/s10462-025-11473-7

Hossain, M. A. (2025). Deep learning-based intrusion detection for IoT networks: A scalable and efficient approach. *EURASIP Journal on Information Security, 2025*, Article 11. https://doi.org/10.1186/s13635-025-00202-w

Karim, F. K., Varela-Aldás, J., Ishak, M. K., Aljarbouh, A., & Mostafa, S. M. (2024). Modeling of Bayesian machine learning with sparrow search algorithm for cyberattack detection in IIoT environment. *Scientific Reports, 14*, Article 28469. https://doi.org/10.1038/s41598-024-79632-4

Keshavamurthy, P., & Kulkarni, S. (2025). An FCM-based hybrid method for DDoS attack detection in resource-constrained devices. *Annals of Telecommunications, 80*, 1071-1094. https://doi.org/10.1007/s12243-025-01130-z

Khan, M. Z., Reshi, A. A., Shafi, S., & Aljubayri, I. (2025). An adaptive hybrid framework for IIoT intrusion detection using neural networks and feature optimization using genetic algorithms. *Discover Sustainability, 6*, Article 382. https://doi.org/10.1007/s43621-025-01141-9

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity, 2*(1), Article 20. https://doi.org/10.1186/s42400-019-0038-7

Korium, M. S., Saber, M., Beattie, A., Narayanan, A., Sahoo, S., & Nardelli, P. H. J. (2024). Intrusion detection system for cyberattacks in the Internet of Vehicles environment. *Ad Hoc Networks, 151*, Article 103330. https://doi.org/10.1016/j.adhoc.2023.103330

Kurniabudi, Stiawan, D., Darmawijoyo, Bin Idris, M. Y., Defit, S., Triana, Y. S., & Budiarto, R. (2022). Improvement of attack detection performance on the Internet of Things with PSO-Search and random forest. *Journal of Computational Science, 61*, Article 101833. https://doi.org/10.1016/j.jocs.2022.101833

Laghrissi, F., Douzi, S., Douzi, K., & Hssina, B. (2021). IDS-attention: An efficient algorithm for intrusion detection systems using attention mechanism. *Journal of Big Data, 8*, Article 119. https://doi.org/10.1186/s40537-021-00544-5

Le Jeune, L., Goedemé, T., & Mentens, N. (2021). Machine learning for misuse-based network intrusion detection: Overview, unified evaluation and feature choice comparison framework. *IEEE Access, 9*, 52312-52325. https://doi.org/10.1109/ACCESS.2021.3075066

Lopez-Martin, M., Carro, B., Sanchez-Esguevillas, A., & Lloret, J. (2017). Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in IoT. *Sensors,*

17(9), Article 1967. https://doi.org/10.3390/s17091967

Luay, M., Layeghy, S., Noorbin, N., Sarhan, M., Kulatilleke, G., Moustafa, N., & Portmann, M. (2026). Time matters: Temporal NetFlow features for ML-based network intrusion detection. *IEEE Access, 14*, 66899-66913. https://doi.org/10.1109/ACCESS.2026.3688204 Advance online publication.

Ma, H., Zhang, W., Zhang, D., & Chen, B. (2025). An IoT intrusion detection framework based on feature selection and large language models fine-tuning. *Scientific Reports, 15*, Article 24085. https://doi.org/10.1038/s41598-025-08905-3

Maksimovic, A. N., Nikolic, V. R., Vidojevic, D. V., Randjelovic, M. D., Djukanovic, S. M., & Randjelovic, D. M. (2024). Using triple modular redundancy for threshold determination in DDoS intrusion detection systems. *IEEE Access, 12*, 53785-53804. https://doi.org/10.1109/ACCESS.2024.3384380

Mananayaka, A. K., & Chung, S. S. (2026). Network intrusion detection with spatial and temporal knowledge graphs and semi-supervised learning. *IEEE Access, 14*, 1071-1094. https://doi.org/10.1109/ACCESS.2026.3706431 Advance online publication.

Megantara, A. A., & Ahmad, T. (2021). A hybrid machine learning method for increasing the performance of network intrusion detection systems. *Journal of Big Data, 8*, Article 101. https://doi.org/10.1186/s40537-021-00531-w

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Pinto, A., Herrera, L.-C., Donoso, Y., & Gutierrez, J. A. (2026). Cyber-physical anomaly detection: A deep adversarial fusion of sensor and network data. *Discover Computing, 29*, Article 70. https://doi.org/10.1007/s10791-026-10064-6

Raza, A., Munir, K., Almutairi, M. S., & Sehar, R. (2023). Novel class probability features for optimizing network attack detection with machine learning. *IEEE Access, 11*, 98724-98739. https://doi.org/10.1109/ACCESS.2023.3313596

Rodríguez, M., Alesanco, Á., Mehavilla, L., & García, J. (2022). Evaluation of machine learning techniques for traffic flow-based intrusion detection. *Sensors, 22*(23), Article 9326. https://doi.org/10.3390/s22239326

Shahzad, F., Mannan, A., Javed, A. R., Almadhor, A. S., Baker, T., & Al-Jumeily, D. (2022). Cloud-based multiclass anomaly detection and categorization using ensemble learning. *Journal of Cloud Computing: Advances, Systems and Applications, 11*, Article 58. https://doi.org/10.1186/s13677-022-00329-y

Singh, A., Chatterjee, K., & Satapathy, S. C. (2021). An edge based hybrid intrusion detection framework for mobile edge computing. *Complex & Intelligent Systems, 8*(5), 3719-3746. https://doi.org/10.1007/s40747-021-00498-4 Advance online publication.

Singh, R., & Ujjwal, R. L. (2023). Hybridized bio-inspired intrusion detection system for Internet of Things. *Frontiers in Big Data, 6*, Article 1081466. https://doi.org/10.3389/fdata.2023.1081466

Sun, W., Li, Z., Cui, J., Sun, L., & Huang, X. (2025). Multimodal spatiotemporal collaborative approach for network traffic anomaly detection. *IEEE Access, 13*, 95955-95968. https://doi.org/10.1109/ACCESS.2025.3598294 Advance online publication.

Tawfik, M. (2024). Optimized intrusion detection in IoT and fog computing using ensemble learning and advanced feature selection. *PLoS ONE, 19*(5), Article e0304082. https://doi.org/10.1371/journal.pone.0304082

Teng, S., Wu, N., Zhu, H., Teng, L., & Zhang, W. (2018). SVM-DT-based adaptive and collaborative intrusion detection. *IEEE/CAA Journal of Automatica Sinica, 5*(1), 204-214. https://doi.org/10.1109/JAS.2017.7510730

Torabi, H., Mirtaheri, S. L., & Greco, S. (2023). Practical autoencoder based anomaly detection by using vector reconstruction error. *Cybersecurity, 6*, Article 2. https://doi.org/10.1186/s42400-022-00134-9

Wang, J., Ge, C., Li, Y., Zhao, H., Fu, Q., Cao, K., & Jung, H. (2025). A two-layer network intrusion detection method incorporating LSTM and stacking ensemble learning. *Computers, Materials & Continua, 83*(3), 5129-5153. https://doi.org/10.32604/cmc.2025.062094

Wang, J., Mishra, R., Singh, S., Sinha, A., Kaur, S., Madhumathi, R., Mishra, S., & Dedhia, P. (2026). A GWO-optimized dual-attention CNN-LSTM model for robust IIoT intrusion detection. *Peer-to-Peer Networking and Applications, 19*, Article 92. https://doi.org/10.1007/s12083-026-02241-7

Wang, L. (2026). MCO-IDM: A network intrusion detection model based on CMO-BOA and Pareto frontier search. *Journal of Cyber Security and Mobility, 15*(3), 149286-149295. https://doi.org/10.13052/jcsm2245-1439.1532 Advance online publication.

Xu, C., Li, D., Liu, Z., Yang, J., Shen, Q., & Tong, N. (2025). Few-shot network intrusion detection method based on multi-domain fusion and cross-attention. *PLoS ONE, 20*(7), Article e0327161. https://doi.org/10.1371/journal.pone.0327161

Yao, W., Hu, L., Hou, Y., & Li, X. (2023). A lightweight intelligent network intrusion detection system using one-class autoencoder and ensemble learning for IoT. *Sensors, 23*(8), Article 4141. https://doi.org/10.3390/s23084141

Yao, W., & Lin, W. (2025). A lightweight anomaly detection model for network traffic using multi-scale spatio-temporal residual learning. *Scientific Reports, 15*, Article 21589. https://doi.org/10.1038/s41598-025-05190-y

Yoshimura, N., Kuzuno, H., Shiraishi, Y., & Morii, M. (2022). DOC-IDS: A deep learning-based method for feature extraction and anomaly detection in network traffic. *Sensors, 22*(12), Article 4405. https://doi.org/10.3390/s22124405

Yousuf, O., & Mir, R. N. (2025). Intelligent network intrusion detection using stacked ensemble of heterogeneous classifiers. *Scientific Reports, 15*, Article 8453. https://doi.org/10.1038/s41598-025-01052-9

Zhang, G., Wang, X., Li, R., Song, Y., He, J., & Lai, J. (2020). Network intrusion detection based on conditional Wasserstein generative adversarial network and cost-sensitive stacked autoencoder. *IEEE Access, 8*, 169048-169060. https://doi.org/10.1109/ACCESS.2020.3031892

Zhang, Z., Wang, P., Zhang, T., Liu, M., & Zhou, X. (2025). Trustworthy generative few-shot learning-based intrusion detection method in Internet of Things. *IEEE Transactions on Consumer Electronics, 71*(1), 1342-1354. https://doi.org/10.1109/TCE.2024.3473304

Zhao, R., Gui, G., Xue, Z., Jiang, J., Ohtsuki, T., Adebisi, B., & Gacanin, H. (2023). A novel intrusion detection method based on lightweight neural network for energy-constrained IoT devices. *Knowledge-Based Systems, 281*, Article 110626. https://doi.org/10.1016/j.knosys.2023.110626

Zheng, L., Wen, Y., Gou, Z., Li, Y., Zhao, M., & Feng, L. (2025). XMF-GNN: An explainable multi-view fusion graph neural network for network intrusion detection. *Neurocomputing, 634*, Article 131285. https://doi.org/10.1016/j.neucom.2025.131285

Zhu, G., Yu, Y., Deng, X., Dai, Y., & Li, Z. (2025). A hybrid split-attention and transformer architecture for high-performance network intrusion detection. *Computer Modeling in Engineering & Sciences, 145*(3), 4317-4348. https://doi.org/10.32604/cmes.2025.074349 Advance online publication.

Zou, L., Luo, X., Zhang, Y., Yang, X., & Wang, X. (2023). HC-DTTSVM: A network intrusion detection method based on decision tree twin support vector machine and hierarchical clustering. *IEEE*

Access, 11, 23169-23184. https://doi.org/10.1109/ACCESS.2023.3251354

APPENDIX

A. CHARACTERISTICS OF INCLUDED STUDIES

Table 6. Characteristics of included studies.

ID	Citation	Primary ML approach	Theme(s)	Quality
S01	Singh et al. (2021)	NB + AdaBoost hybrid (EHIDF); C4.5 signature module	T1, T4	Adequate
S02	Megantara and Ahmad (2021)	DT + RFE; LOF anomaly detection	T1	Adequate
S03	Shahzad et al. (2022)	1D-CNN + LSTM; SGD/LR/Ridge voting ensemble	T3, T4	Adequate
S04	Laghrissi et al. (2021)	LSTM + Attention; PCA/MI feature reduction	T3	Adequate
S05	Karim et al. (2024)	Bayesian Belief Network; COA + SSA optimisation	T1	Adequate
S06	Elsaid et al. (2024)	GRU + LSTM; GWO feature selection	T3	Adequate
S07	Altamimi and Abu Al-Haija (2024)	Extreme Learning Machine (ELM); SMOTE	T1	Adequate
S08	Torabi et al. (2023)	Autoencoder; vector reconstruction error threshold.	T3	Adequate
S09	Ahmed et al. (2024)	Stacking ensemble (DT/RF/MLP/LGBM); BMA;	T2	Adequate
S10	Elsedimy et al. (2024)	CEM QSVM + IGWO; hybrid signature + anomaly detection	T1, T7	Adequate
S11	Wang et al. (2026)	CNN-LSTM + dual attention (spatial + temporal)	T3	Adequate
S12	Francis et al. (2026)	Survey of swarm-intelligence feature selection for	T5	Adequate
S13	Yao and Lin (2025)	IIoT IDS Multi-scale 1D-CNN; knowledge distillation	T3, T6	Adequate
S14	Keshavamurthy and Kulkarni (2025)	Fuzzy Cognitive Maps; SelectKBest feature scoring	T1	Adequate
S15	Khan et al. (2025)	ANN + GA/PSO optimisation; FW-SMOTE	T1	Weak
S16	Ma et al. (2025)	RF + GPT-2 data augmentation; MCP feature	T1, T5	Strong
S17	Hossain (2025)	Selection 1D-CNN vs LSTM; multi-dataset comparison	T3	Adequate
S18	Pinto et al. (2026)	VAE-LSTM + AAE; multimodal CPS (sensor + network)	T3, T8	Strong
S19	Teng et al. (2018)	Integer programming classifier fusion	T1	Adequate
S20	Ahmed et al. (2020)	Online RF; adaptive streaming detection	T1, T6	Adequate

S21	Le Jeune et al. (2021)	RF + SVM + DT ensemble; real traffic evaluation	T1	Strong
S22	Zhang et al. (2020)	CWGAN + cost-sensitive RF; minority	T3, T5	Adequate
S23	Maksimovic et al. (2024)	Oversampling AdaBoost + ITRS over-voting ensemble	T2	Adequate
S24	Zou et al. (2023)	Hierarchical clustering + ELM (HC-FIDS)	T1, T4	Adequate
S25	Raza et al. (2023)	Two-stage CNN-BiLSTM	T3, T4	Adequate
S26	Abdulboriy and Shin (2024)	Streaming SVM/RF/DT ensemble; online learning	T1, T6	Adequate
S27	Luay et al. (2026)	Transformer with NF3 normalisation flow; multi-dataset	T3	Strong
S28	Eltanbouly et al. (2025)	Survey of ML, DL, and GNN techniques for IDS	T1, T3, T9	Adequate
S29	Mananayaka and Chung (2026)	Knowledge graph + GNN; semi-supervised detection	T3, T9	Adequate
S30	Sun et al. (2025)	CNN-BiLSTM + self-attention; SMOTE	T3	Adequate
S31	Zhang et al. (2025)	CDDPM data augmentation + CNN; few-shot learning	T3, T5	Adequate
S32	Wang (2026)	BOA-optimised SVM	T1	Adequate
S33	Lopez-Martin et al. (2017)	CVAE; missing-data imputation for IDS	T3	Adequate
S34	Yoshimura et al. (2022)	CNN + threshold calibration; semi-supervised anomaly	T3, T6	Adequate
S35	Rodríguez et al. (2022)	SVM, RF, k-NN, DT comparative study	T1	Adequate
S36	Tawfik (2024)	RF + SVM ensemble; fog node deployment	T1, T6	Adequate
S37	Yao et al. (2023)	One-class autoencoder (OCAE); N-BaIoT edge IDS	T3, T6	Adequate
S38	Singh and Ujjwal (2023)	SVM + ACO/GA hybrid feature selection	T1	Adequate
S39	Xu et al. (2025)	Few-shot CNN + bidirectional cross-attention;	T3	Adequate
S40	Alqahtani et al. (2025)	2D-DCT RF, SVM, k-NN, LR, NB comparative study;	T1	Adequate
S41	Yousuf and Mir (2025)	SMOTE LightGBM stacking ensemble; SMOTE	T2	Adequate
S42	Alshammari et al. (2026)	1D-CNN + LSTM + GRU deep ensemble	T3	Adequate
S43	Al-Daweri et al. (2025)	LSTM + ADASYN; minority class oversampling	T3, T5	Adequate
S44	Kurniabudi et al. (2022)	Study PSO-Search grouped feature selection + RF	T1	Adequate
S45	Korium et al. (2024)	Regression-based feature	T1	Adequate

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324

		selection; SVM, RF, k-NN		
S46	Zhao et al. (2023)	OC-SVM + MCU edge deployment; NF-v2 NetFlow	T1, T6	Strong
S47	Aljameel et al. (2025)	DDPM diffusion augmentation + RF downstream	T3, T5	Adequate
S48	Zheng et al. (2025)	GNN + knowledge graph; graph-based IDS	T3, T9	Adequate
S49	Zhu et al. (2025)	WGAN + CNN; adversarial data augmentation	T3, T5	Adequate
S50	Wang et al. (2025)	Hierarchical RF + NMRI feature importance	T1, T4	Adequate

B. DATA EXTRACTION FIELDS

Table 7. Data extraction fields applied uniformly to all 50 included studies.

Field	Description
Bibliographic details	Authors, year, title, journal or conference, volume, pages, DOI.
Study environment	Network type (enterprise LAN, IoT, ICS/CPS, cloud, mixed); traffic protocol (TCP/IP, NetFlow, PCAP, other).
ML technique family	Classification into one or more of nine synthesis themes (T1-T9) based on the primary detection method.
Specific algorithms	Named ML or DL algorithms reported, including feature selection and preprocessing methods.
Evaluation dataset(s)	Dataset name, size (records and features), attack categories covered, and public availability.
Performance metrics	All quantitative metrics reported: accuracy, precision, recall, F1-score, area under the curve, false positive rate, Matthews correlation coefficient, and any others.
Evaluation protocol	Train/test split ratio or cross-validation scheme; class balance or imbalance handling; random seed reported.
Deployment indicators	Evidence of real time operation, edge or resource-constrained deployment, inference latency, or model size reporting.
Quality flags	Study specific methodological concerns flagged during extraction (see Section 2.6).

C. QUALITY ASSESSMENT CRITERIA

Table 8. Quality assessment criteria applied to all included studies.

Rating	Criteria
Strong	≥3 independent datasets; re-implemented or directly comparable baselines; per class F1 for minority attack classes; ≥1 efficiency indicator (latency, model size, or throughput); no irreconcilable methodological inconsistencies.
Adequate	Sound design with ≤1 of: single dataset; published-figure baselines only; minority class performance unreported; efficiency unmeasured; statistical significance untested. Primary claims supported by reported evidence.
Weak	≥1 of: irreconcilable dataset/results inconsistencies; no quantitative baseline on same dataset;

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23385324