Digital Innovations & Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 14. No. 3, September 2026 Series

Article Citation Format
Abdullahi, M.B., Ndanusah, U.H., Subairu, S.O., Ahmad, S. & Noel, M.D. (2026): Knowledge Distillation for Lightweight Network Intrusion Detection: A Systematic Literature Review of Single and Multi-Teacher Approaches.. Journal of Digital Innovations & Contemporary Research in Science, Engineering & Technology.
Vol. 14, No. 3. Pp 31-47. www.isteams.net/digitaljournal
dx.doi.org/10.22624/AIMS/DIGITAL/V14N3P3

Article Progress Time Stamps
Article Type: Research Article
Manuscript Received: 12th July, 2026
Review Type: Blind Peer
Final Acceptance: 2nd September, 2026

Knowledge Distillation for Lightweight Network Intrusion Detection: A Systematic Literature Review of Single and Multi-Teacher Approaches.

[2]Abdullahi, M.B., [1]Ndanusah, H.U., [1]Subairu, S.O., [3]Ahmad, S. & [1]Noel, M.D.
[1]Department of Cyber Security Science, Federal University of Technology, Minna, Nigeria
[2]Department of Data Science, Federal University of Technology, Minna, Nigeria
[3]Department of Intelligence and Security Science, Federal University of Technology, Minna, Nigeria
E-mails: [2]el.bashir02@futminna.edu.ng, [1]halirundanusah@gmail.com,
[1]sikiru.subairu@futminna.edu.ng, [3]ahmads@futminna.edu.ng, [1]moses.noel@futminna.edu.ng

ABSTRACT

The proliferation of Internet of Things (IoT) infrastructures, software-defined networks, industrial platforms and edge computing has surged the need for network intrusion detection systems (NIDSs) which can maintain high detection accuracy while working with strict memory, latency and energy constraints. Deep learning models can recognise challenging traffic patterns, but their computational strain frequently limits direct use on smart constrained devices. Knowledge distillation (KD) has been a viable compression technique that transfers supervisory knowledge between a big teacher and a tiny student. This review is scoped as a systematic review of KD for lightweight NIDS, and empowers EKD as an emerging multi-teacher paradigm for analysis. Following PRISMA 2020, five databases were searched to find studies published from January 2018 to 31 March 2026. After screening a total of 847 records, forty-two (42) met the eligibility criteria and were included. The dominant proportion of the corpus comprised response-based distillation approaches (32 studies, 76%), followed by feature-based distillation approaches (7 studies, 17%), then hybrid response-feature approaches (2 studies, 5%) and relation-based approaches (1 study, 2%). Of the total, only 13 studies (31%) tested some type of physical edge device, while only 9 (21%) included adversarial robustness testing (only 3 of which tested both types of devices). From the synthesis, it is observed that output-level distillation is still popular since it is architecture-agnostic, is cheap to implement and simple to stack on top of teachers, while feature-based transfer necessitates alignment from layer, projection functions and demands high memory during training. For NIDS, ensemble supervision has substantial security justification compared to many customary identification tasks because class imbalance and temporal dependence are both usually present in traffic data, along with protocol diversity, and the possibility of attacks to manipulate protocols can be considered a rare circumstance. Hence, heterogeneous teachers, hardware-aware benchmarking, explicit loss-function reporting and robustness testing are considered to be major avenues for future research in the review.

Keywords: Network Intrusion Detection, Lightweight Deep Learning, Edge Computing, Traffic, Data Ensemble Knowledge Distillation, *, Knowledge Distillation, Adversarial Robustness.*

1. INTRODUCTION

As the malicious traffic proliferates through IoT systems, industrial control networks and software-defined infrastructure, cloud services and distributed edge, network intrusion detection is a core aspect of cyber defence. Non-linear traffic patterns can be learned by deep neural models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory networks (LSTMs) and Transformers, that many conventional classifiers handle without optimal result. However, they may have considerable parameters counts, memory requirement and inference cost to get its detection performance. These requirements have maintained the curiosity for lightweight NIDS models that can operate in close proximity to the traffic source without always depending on the remote servers [1],[2],[3].

Knowledge distillation tackles this issue by training a compact student model from a larger teacher so it mimics information provided by the teacher. While in the classical response-based KD the student learns from the teacher's soft class probabilities that highlight inter-class relationships not represented in hard labels [4]. Feature-based methods also include transferring to intermediate encoder representations and may often need the help of a projection function if teacher and student dimensions are different [5]. Relation-based approaches, on the other hand, transfer pairwise distances, angles or other relational structures between representations as this can permit transferring without forcing direct equality of features [6]. These alternatives bring different compromises in the number of transfers vs. burden of implementation and the compatibility across the different families of models.

It is important for the scope of this review to note the difference between single-teacher KD and ensemble knowledge distillation. A total of 40 KD based NIDS studies have been supplied, 30 of which use single teacher. 12 have multiple teachers (8 homogeneous ensembles, 4 heterogeneous ensembles). Hence, it would be an overstatement to refer to ensemble supervision as “reaching maturity” if all 42 papers would be considered. This review sets general KD in lightweight NIDS as the main field and examines EKD as a secondary but developing subset. The framing of the gap in the literature, eligibility criteria, research questions and conclusions is with the evidence that was actually synthesised.

NIDS also create security specific reasons to investigate ensemble supervision. Samples in image-recognition problems tend to be spatially structured, while classes are reasonably stable, while attack classes may be highly imbalanced, local structures in the network traffic be present, and long temporal sequences be present, and, finally, network protocols may change over time, and attacks flow may be manipulated adversarially. A CNN teacher could be used to capture local interactions between features, an LSTM could model temporal dynamics, and a Transformer could capture longer range relations. With a heterogeneous teacher, students can thus be exposed to complementary supervisory signals. This diversity is particularly crucial when the scarce sample of existing labelled examples of rare attacks is insufficient with one architecture or when one architecture overfits to a specific type of traffic. While ensemble is not a guarantee for better detection or robustness, it does show a principled way to limit overly relying on one teacher's mistakes [7], [8].

The performance of lightweight should be judged based on security deployment. Clean-data accuracy is beyond proof of readiness. The way a deployable NIDS should present its report should include the following: number of parameters, size of model, latency, throughput, memory, numerical precision, and energy usage and robustness when faced with adversarial manipulation. Distillation reducing the model size and inference demand does not automatically translate to adversarial protection. Defensive distillation was suggested as a defence mechanism, but it was found that the attacks can be made in such a way so as to circumvent the defensive distillation [9], [10]. FGSM and PGD are still important baseline attacks for direct robustness evaluation [11], [12]). Based on this background, the review builds and synthesizes Teacher architectures, distillation losses, deployment evidence and robustness evaluation. It goes beyond frequency counts and examines the mathematical nature and implications of the distinction between a response, feature, relation and hybrid distillation, and neatly untangles the difference between model compression and actual performance on physical edge devices.

1.1 Main Research Objectives

The main objective of this work is to collate existing know-how on knowledge distillation for lightweight NIDS models, and figure out the effect of ensemble supervision on the design and assessment of compact NIDS models.

1.1.1 Specific Research Objectives

1. Classification of teacher architectures designs as either of single, homogeneous ensemble or as a heterogeneous ensemble.
2. Comparison of response, feature, relation and hybrid distillation strategies and corresponding loss functions
3. Analysis of where and how deployment constraints (including parameter count, model size, latency, memory, numerical precision, and energy) are measured in the studies.
4. Analysis of how frequent a distilled NIDS models received direct adversarial robustness testing.

1.2. Research Questions

General Research Question

How is knowledge distillation, especially rising ensemble knowledge distillation techniques is been designed, evaluated, and applied to growing lightweight, efficient, and robust network intrusion detection systems, and what are the evidence for the suitability of knowledge distillation for application in a resource-constrained environment?

Specific Research Questions

- **RQ1:** What teacher architectures and teacher multiplicity configurations have been investigated in lightweight NIDS using knowledge distillation?
- **RQ2:** What are the knowledge distillation strategies and loss functions used in the lightweight NIDS studies, and what are the considerations and performance trade-offs of response-based, feature-based, relation-based, and hybrid distillation approaches?
- **RQ3:** What is the rigor of existing studies for the edge deployment requirements, including the model parameters, model size, target hardware platforms, numerical precision, inference latency, memory, throughput, and energy consumption?
- **RQ4:** How far do existing research studies evaluate the adversarial robustness of distilled NIDS models in realistic attack scenarios like Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD)?

1.3 Methodology

The review was conducted in accordance with the PRISMA 2020 reporting principles by structuring the identification, screening, eligibility assessment and synthesis [13]. The search concepts and inclusion/exclusion criteria, extraction fields and themes for synthesis, were determined in advance before concluding the final screening. The protocol was not publicly registered, which is a limitation as the external readers have no option of judging if there were any alteration of the decision after the screening initiation.

2. LITERATURE SOURCES AND SEARCH STRATEGY

Searches were conducted in Scopus, Web of Science, IEEE Xplore, ACM Digital Library and arXiv. The review period refers to the period 1st January, 2018 to 31st March, 2026. To gain broad multidisciplinary indexing, articles were drawn from Scopus and Web of Science, as well as from IEEE Xplore and ACM Digital Library that highlight the literature of computing and cyber-security conferences which may otherwise be absent in general databases. arXiv was checked for recent work, but if a preprint was identical to a peer-reviewed paper the peer-reviewed work was preferred.
The search query consisted of 3 concept blocks: knowledge distillation, network intrusion detection and lightweight/edge deployment. The synonyms within each block were joined by OR and the blocks were joined together by AND. The robustness terminology was used separately in a robustness search so as not to filter out otherwise eligible lightweight KD studies not involving adversarial testing**Search Keywords and Search Terms**

Concept 1 (knowledge distillation):
“Knowledge distillation”, “ensemble knowledge distillation”, “multi-teacher knowledge distillation”, “multi-teacher distillation”, “multiple-teacher distillation”, “dual-teacher distillation”, “teacher aggregation”, “teacher ensemble”, and “teacher-student learning”.

Concept 2 (network intrusion detection):
“Network intrusion detection”, “network intrusion detection system”, “intrusion detection system”, NIDS, “network anomaly detection”, “malicious traffic detection”, and “network attack detection”.

Concept 3 (lightweight or edge deployment):
lightweight, “light-weight”, compact, “lightweight model”, “compact model”, “model compression”, “resource-constrained”, “edge computing”, “edge device”, IoT, “Internet of Things”, embedded, latency, “memory footprint”, “energy consumption”, throughput, and “inference time”.

Master Boolean search string:
((“knowledge distillation” OR “ensemble knowledge distillation” OR “multi-teacher knowledge distillation” OR “multi-teacher distillation” OR “multiple-teacher distillation” OR “dual-teacher distillation” OR “teacher aggregation” OR “teacher ensemble” OR “teacher-student learning”) AND (“network intrusion detection” OR “network intrusion detection system*” OR “intrusion detection system*” OR NIDS OR “network anomaly detection” OR “malicious traffic detection” OR “network attack detection”) AND (lightweight OR “light-weight” OR “lightweight model*” OR compact OR “model compression” OR “resource-constrained” OR “edge computing” OR “edge device*” OR IoT OR “Internet of Things” OR embedded OR latency OR “memory footprint” OR “energy consumption” OR throughput OR “inference time”))

Supplementary adversarial-robustness search string:
((“knowledge distillation” OR “ensemble knowledge distillation” OR “multi-teacher distillation” OR “teacher aggregation”) AND (“network intrusion detection” OR “intrusion detection system*” OR NIDS OR “network anomaly detection”) AND (“adversarial robustness” OR “adversarial attack*” OR “adversarial example*” OR FGSM OR “Fast Gradient Sign Method” OR PGD OR “Projected Gradient Descent” OR “Carlini-Wagner” OR “robust accuracy” OR “attack success rate”))

2.1 Eligibility Criteria

The selected studies must propose or evaluate deep-learning knowledge distillation for network intrusion detection, provide quantitative results on the detection performance, and incorporate at least a measure of compression, computational efficiency or deployment performance. Journal papers as well as conference papers were acceptable. Review articles have been used only to find citation and background; they are not included in the main synthesis.

Table 1. Updated the eligibility criteria for KD based Light weight NIDS review.

Criterion	Inclusion	Exclusion
Population/context	NIDS based on network-traffic data in enterprise, IoT, SDN, cloud, vehicular or edge networks.	Host-based detection, malware-only classification, or non-network security tasks.
Intervention	Deep-learning knowledge distillation for a lightweight NIDS, including single-teacher, multi-teacher, dual-teacher, teacher aggregation, self-distillation or ensemble distillation.	NIDS without KD; pruning, quantisation or feature selection without a distillation component.
Comparator	Teacher model or ensemble, undistilled student, uncompressed model, conventional NIDS, or another compression baseline.	Claims that cannot be linked to an interpretable numerical baseline.
Outcomes	Quantitative detection performance plus at least one lightweight indicator such as parameters, model size, FLOPs, latency, memory, energy, throughput or physical edge evaluation. Robustness extracted when reported.	No empirical assessment or no evidence related to lightweight deployment.
Study design	Full empirical journal or conference article; eligible arXiv version only where no duplicate peer-reviewed version existed.	Reviews, surveys, editorials, protocols, posters, extended abstracts, theses, blogs, duplicates or inaccessible reports.
Language	English.	Non-English.
Time frame	1 January 2018 to 31 March 2026.	Outside the review period.

2.2 Study Selection, Data Extraction, Snowballing and Methodological Appraisal

The study selection process was implemented in two phases: title and abstract screening and full-text evaluation. Extraction fields included publication type, dataset, classification, teacher architecture, student architecture, teacher multiplicity, distillation strategy, loss, temperature, loss-weight (if reported), parameter count, model size, target hardware, numerical precision, inference framework, latency, throughput, memory, energy and adversarial robustness.

This extended extraction scheme lets the extraction separate what happens during training time from what happens during deployment time and lets the review compare models that claim to be 'lightweight' in different ways.

2.3 Backward and Forward Snowballing

In backward snowballing, the reference lists of full-text eligible primary studies and relevant review papers that served as discovery aids were examined.

Table 2. Canvass of the KD-based NIDS studies as per the MMAT.

Appraisal domain	Operational rating rule
Q1. Research question and design	Yes: research aim, NIDS task and KD objective are explicit. Partial: one element is vague. No: objective or design cannot be determined.
Q2. Dataset appropriateness and provenance	Yes: dataset source, classes, preprocessing and split strategy are reported and suitable. Partial: one or more details are incomplete. No: dataset use cannot be assessed.
Q3. Comparator and experimental control	Yes: teacher/student and at least one interpretable baseline use comparable data and evaluation conditions. Partial: baseline is incomplete or conditions differ. No: no valid comparator.
Q4. Outcome validity and statistical reporting	Yes: suitable detection metrics are reported with repeated runs, uncertainty or statistical testing where appropriate. Partial: metrics are adequate but uncertainty is absent. No: outcome reporting is inadequate.
Q5. Distillation reproducibility	Yes: teacher/student architectures, loss terms, temperature, weights and training settings are sufficient to reproduce KD. Partial: some settings are missing. No: KD implementation cannot be reconstructed.
Q6. Edge-deployment validity	Yes: physical target hardware and deployment settings are reported with latency plus memory, size, throughput or energy. Partial: only a subset is reported or hardware is simulated. No: deployment claims rely only on desktop inference or model size.
Q7. Numerical precision and optimisation reporting	Yes: FP32/FP16/INT8 or other precision and any pruning/quantisation settings are explicit. Partial: optimisation is named but precision details are missing. No: numerical format is unreported where it affects deployment claims.
Q8. Security robustness	Yes: adversarial or distribution-shift evaluation uses clearly specified attacks, budgets and security metrics. Partial: limited robustness test. No: no robustness assessment.
Q9. Reproducibility and transparency	Yes: code/configuration or sufficient implementation detail is available. Partial: some information is available. No: replication is materially obstructed by missing detail.

Forward snowballing then checked papers that cited these seed records through Scopus and Web of Science, with Google Scholar used as a supplementary cited by source when a database record did not give adequate citation analysis. Chasing candidate records based on citations were then deduplicated with the result of the database search, and screened following the same criteria date, language, intervention, NIDS and lightweight-outcome, as the main search. Record lists of candidates could be produced from review papers but these were not included in the primary study synthesis. The final date for citations to continue was 31 March 2026.

2.4 Data Synthesis Strategy

The datasets were different, and attack taxonomies differed, as well as teacher-student architectures, efficiency measures, and robustness settings, rendering statistical meta-analysis inappropriate. Review is therefore performed utilising the technique of structured narrative synthesis. Descriptive frequencies provide for the distribution of teacher and distillation designs and analytical taxonomy provides for representations of comparisons of teacher and distillation designs in respect to loss functions, teacher-compatibility, training overhead and likely deployment implications. Only when the study provides sufficient detail to relate the performance to an execution environment will hardware evidence be used. This means the approach will not sacrifice small model file size or rapid desktop inferences for the sake of being "edge ready".

3. RESULTS

847 records were found in the search. Following removal of 203 duplicates, 644 records went to title- and abstract screening. Of these, 580 did not meet the inclusion criteria, leaving 64 full-text reports. Twenty-two full text reports did not meet the criteria for empirical and lightweight or NIDS eligibility and 42 studies were included in the final qualitative synthesis.

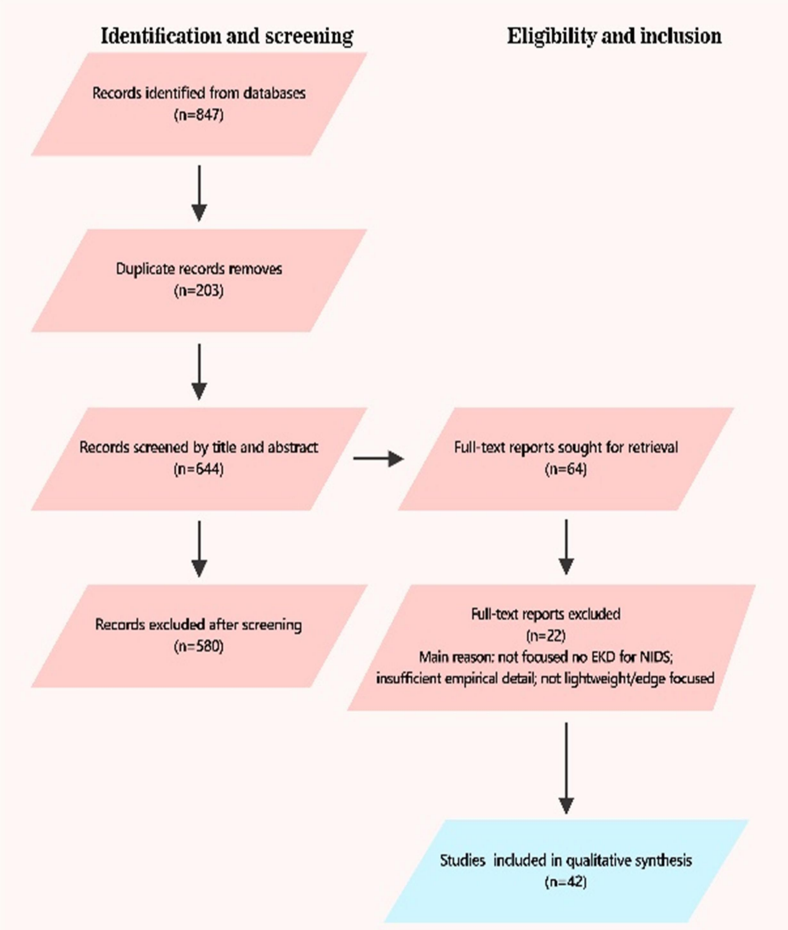

Figure 1. Flow chart of the process for study selection (PRISMA 2020)

3.1 Study Characteristics

A final volume of 28 journal articles and 14 conference papers was included in the final corpus. 22 studies were published from 2024 to March 2026, compared with 15 from 2021 to 2023 and 5 from 2018 to 2020. The most frequent data used was CICIDS2017, followed by NSL-KDD and UNSW-NB15. The evidence base was dominated by binary classification, with 33 studies using binary classification and nine using multiclass classification.

Table 3. characteristics of included studies (N = 42).

Characteristic	Frequency	Percentage
Journal articles	28	67%
Conference papers	14	33%
2018–2020	5	12%
2021–2023	15	36%
2024–2026	22	52%
CICIDS2017	23	54%
NSL-KDD	10	24%
UNSW-NB15	5	12%
Binary classification	33	79%
Multiclass classification	9	21%

RQ1: Teacher Architectures and the actual Scope of KD

Teacher architecture configuration shows that EKD is a minority of the evidence base. There were 30 studies that employed a single teacher (71% of the studies). Twelve studies involved several teachers, eight homogeneous and four heterogeneous. Hence, evidence indicates that there is room for a KD in the context of lightweight NIDS where EKD would be a developing branch, rather than a description of the whole corpus as ensemble distillation.

Table 4: Teacher architecture distribution and scope implication

Teacher architecture	Studies	Percentage	Interpretation
Single teacher	30	71%	General KD baseline; no ensemble supervision.
Homogeneous ensemble	8	19%	Multiple teachers from the same broad model family or repeated variants.
Heterogeneous ensemble	4	10%	Different model families, for example CNN combined with LSTM or Transformer teachers.

The small heterogeneous subgroup is important because there are a number of forms of structure in the NIDS traffic. CNNs model local interactions between flow features, recurrent models model time dependence and Transformers model longer-range associations. The learning signals a student receiving from these various teachers can be more diverse than the learning signals a student receiving from one architecture. But having different teachers, at the same time introduces issues of calibration and differing predictions and reliability among the attack classes. Therefore, explicit aggregation rules like uniform averaging, confidence-weighted aggregation, or uncertainty-aware aggregation should be adopted in multi-teacher systems as opposed to considering it as a design choice.

RQ2: Distillation Strategies, Loss Functions and Trade-Offs

Among 42 studies, 32 (76%) used response-based distillation, 7 (17%) used feature-based transfer, 1 (2%) used relation-based transfer, and 2 (5%) used hybrid response feature transfer. These frequencies do not refer to comparative superiority, but rather to adoption. The advantage of response-based KD is that it does not necessitate matching dimensions in the hidden layers and that it can be readily implemented at different level across various teacher/student architectures. The feature-based KD can convey more internal information, but it depends on the researchers to determine the internal correspondence content of a layer and (usually) train a projection layer. Stating dimensional equality is important for a KD such as the relation-based KD, which, instead of assuming dimensional equality, matches structural relations between examples, but then depends on pairwise or higher order calculations that can lead to higher training costs. These signals can be paired with a variety of hybrid methods to add more loss weights and tuning decisions.

Table 5. Distribution of distillation techniques.

Distillation strategy	Studies	Percentage
Response-based	32	76%
Feature-based	7	17%
Relation-based	1	2%
Hybrid response + feature	2	5%

A classical response-based objective functions are based on the supervised cross-entropy with a parameter tuned teacher-student divergence. For the teacher, student logits z_t and z_s, while temperature be denoted by T, $p_t^T = softmax(z_t/T)$, $p_s^T = softmax(z_s/T)$, and α the weight on the distillation term α. Common formulation are: $L = (1-\alpha)L_{CE}(y, p_s) + \alpha T^2 D_{KL}(p_t^T || p_s^T)$. (Equation 1)

Ensemble response distillation can first make weighted teacher distribution for M teachers.

If $w_m \geq 0 \; and \; \sum_m w_m = 1, then: \; p_{T=\sum_m w_m}^T \; p_{(t_m)}^T, L_{MTKD} = T^2 \; D_{KL(p_t^T || p_s^T)}$. (Equation 2)

Uniform weights implement simple averaging. Confidence or uncertainty-based weights can give more influence to teachers that are reliable for a particular flow or attack class, but the review shows that adaptive weighting remains uncommon. Feature-based distillation matches an internal teacher representation h_t with a projected student representation $P(h_s)$. The projection is needed when teacher and student dimensions differ:

$L_{hint} = \| P(h_s) - h_t \|_2^2$. (Equation 3)

A student may have multiple teachers, each one has a set of projected targets, in which case, a student could match multiple projected targets, for instance, $L_{MF} = \sum_m w_m \| p_m (h_s) - h (t_m) \|_2^2$. This formulation might carry more enriched hidden information than logit matching, when each extra teacher might need one extra alignment and could cost one extra memory for training. For Relation-based KD, it is relationships between the teacher and the student representations that are matched, e.g. pairwise distances or angles: $L_{rel} = \| R(H_s) - R(H_t) \|_F^2$, where $R(.)$ stands for some relational operators acting on a batch of examples.

When the teacher and student hidden spaces are different dimensionalities, this can be beneficial as geometry will be transferred rather than raw feature vectors [6].

Table 6. Analytical taxonomy of distillation losses, architectures and trade-offs.

Strategy	Typical loss	Cross-architecture compatibility	Training burden	Potential NIDS benefit	Main limitation
Response / logit matching	KL divergence or cross-entropy between softened teacher and student outputs.	High; teacher and student hidden dimensions can differ.	Low to moderate.	Easy to aggregate across multiple teachers; preserves class-confusion information; suitable for heterogeneous teachers.	May discard useful intermediate structure; teacher miscalibration can be transferred.
Feature / hint matching	L2, cosine or other distance between teacher features and projected student features.	Moderate; requires layer selection and projection when dimensions differ.	Moderate to high.	Can transfer internal representations relevant to rare attacks or subtle feature interactions.	Layer alignment is difficult across CNN, LSTM and Transformer teachers; additional training memory and tuning.
Relation-based	Distance-, angle- or similarity-preserving losses between examples or feature relations.	Moderate to high across differing feature dimensions.	Moderate to high, especially for pairwise relations.	Can preserve structural separation between normal and attack traffic without exact feature equality.	Sparse NIDS evidence; cost can rise with batch size; relation definition may be task-sensitive.
Hybrid	Weighted sum of response, feature and/or relation losses.	Depends on components.	High.	Can combine class-level and representation-level supervision.	Requires several hyperparameters and careful ablation to show which signal produces gains.

It is therefore not surprising that the response-based KD dominates 76% of the results for three reasons. Firstly, there isn't a need to expose and align internal layers; the logits are accessible at model output. Second, the response-level aggregation is straight-forward when the teachers are heterogeneous because each teacher generates his/her class probabilities even if the hidden representation is not the same.

Third, NIDS studies try to minimise the engineering overhead on deploying these systems, and minimise the volume of the systems themselves, so an approach that modifies the training objective without requiring any additional inference modules on the clients is appealing. The above statements need not be interpreted as evidence that logits are always better. Low feature and relation losses can also be partially due to reporting and implementing limitations, rather than low performance.

RQ3: Edge Deployment and Hardware Analysis

The evaluation in the real world is still in its infancy. Only 13 (31%) of the 42 studies tested physical edge devices, like Raspberry Pi or NVIDIA Jetson devices, being tested. When reported, model sizes ranged from 20KB to 500KB, and inference latency ranged from 0.07 to 2.1ms. But the amount of energy consumed was rarely reported and the reporting since ended up incomplete regarding specification of the inference framework, the numerical precision and the memory footprint. All of these omissions make the comparison of lightness claim or readiness to deploy difficult among studies.

Table 7. hardware interpretation of the evidences for deployment on the edge.

Hardware dimension	Evidence currently available	Comparability problem	Required reporting standard
Physical edge evaluation	13 studies (31%)	29 studies (69%) did not test physical edge hardware.	Physical validation should be separated from desktop or simulated claims.
Target hardware	Raspberry Pi and NVIDIA Jetson devices are mentioned across the edge-tested subset.	Counts by exact hardware model are not preserved in the supplied extraction.	Report exact board, CPU/GPU, RAM, power mode and operating system.
Model parameters	Not consistently retained in the aggregate table.	Cannot relate model size to architectural scale.	Report trainable parameters and, where relevant, FLOPs/MACs.
Model size	20 KB–500 KB pooled range.	Range is not linked to hardware or numerical precision.	Report on-disk and in-memory size with precision format.
Numerical precision / quantisation	Inconsistently reported.	FP32, FP16 and INT8 models can have very different memory and latency.	State numerical precision, quantisation method and calibration procedure.
Inference latency	0.07–2.1 ms pooled range.	Batch size, framework and warm-up procedure are often absent.	Report median and variability, batch size, framework and whether preprocessing is included.
Memory footprint	Inconsistently reported.	Model file size does not equal peak runtime memory.	Report peak RAM/VRAM during inference.

RQ4: Adversarial Robustness

A total of 33 (79%) studies did not assess adversarial robustness. The most popular attacks reported were FGSM and PGD. None of the study comes up with a standardised baseline, combining KD with adversarial training under the comparable white and black-box setting.

This means that this is important for a lightweight NIDS if an attacker tries to modify traffic properties or takes advantage of the weakness of its model. Thus, use of distillation should initially be considered as a compression technique and any robustness should be argued based on adversarial testing rather than the assumption of smoother outputs.

Table 8. The adversarial robustness reporting in the studies reported

Characteristic	Frequency	Percentage
Adversarial robustness evaluated	9	21%
No adversarial robustness evaluation	33	79%
FGSM reported	9	21%
PGD reported	9	21%
Combined KD + adversarial training as a common baseline	0	0%

Methodological Quality and Reporting Transparency:

The appraisal results showed good performance for clear objectives & description of dataset, but low performance in edge deployment, adversarial robustness & reproducibility. Methodological quality of the included research was generally good (6.2/9). The studies mostly had defined objectives and an adequate description of datasets. In contrast, the ability to deploy to the edge and adversarial robustness testing were the least strong dimensions. Reproducibility was also not consistent across studies as many were not reporting enough information on how to replicate the implementation. The quality profile of the evidence base is summarised in Table 9.

Table 9: Aggregate quality assessment of studies included.

Legacy quality indicator	High	Moderate	Low
Clear research objectives	78%	17%	5%
Dataset and data reporting	83%	14%	3%
Methodological clarity	64%	24%	12%
Edge deployment evaluation	12%	19%	69%
Adversarial robustness testing	8%	13%	79%
Reproducibility of results	55%	30%	15%

4. DISCUSSION

4.1 Why Ensemble Supervision has a Security-Specific Rationale

Many signal types and asymmetric error costs exist in network traffic; ensemble supervision can be very beneficial in NIDS. These effects can result in extreme class imbalance in the case of rare attacks, attacks over time can take place across sequences of flows, and the distribution of features can change between different protocols or different devices. One teacher might dominate attacks for one traffic while underrepresenting attacks for a minority traffic. The variety of models can be integrated using a heterogeneous ensemble, so that the student could be then complemented with local-pattern and temporal or long-range models.

From an operational perspective, this is significant because false negative results can put a system at risk, and false positive results can cause an overload of analysts or automated response systems. Ensemble distillation therefore has a robust justification when it can optimize minority-class recall, calibration or stability with no cost at deployment. But there are new failure modes with multi-teacher KD. All teachers are subject to training bias, systematic disagreement with other rare classes or poorly calibrated confidence. Simple average can minimize a specialist teacher that is quite good at one attack family. To conclude, future NIDS studies should measure the non-uniform averaging against a confidence-aware/user uncertainty-aware weighting, and also present per-class effects, particularly for those attacks that have lower frequencies. Rather, the diversity of teachers should be measured and analyzed rather than taken for granted from architecture names alone.

4.2 Why Response Based Distillation is the Dominant One

The response-based KD at 76% is due to the convenience of the implementation and cross architectures. Logits are used across CNNs, LSTMs, Transformers and MLPs. For EKD, it is important that the size of the teacher set can be heterogeneous, and that the alignment of these features could be challenging because of the possibly different dimensions of the layers, sequence length, and internal representations with which these teachers operate.

The output probabilities avoid these choices for alignment. In addition, student inference graph remains the same with the Response-based KD as the student is aggregated by the teacher at training time with an additional temperature scaling. For NIDS studies focused on an edge device, that simplicity reduces risk that a compression method introduces deployment overhead.

Methods that are based on features constitute only 17% of the corpus partly because they involve more design decisions. For the researchers, they need to choose their teacher and student layers, specify their projection functions and then choose if the loss should be direct with the activations or with the normalised features/attention maps. A weak relationship between representations may occur during CNN to LSTM or Transformer to MLP transfer.

Relations between dimensions, such as matching distances, similarities, can reduce this dimensional mismatch, which is only represented in the present corpus by one relation-based study, thus this situation doesn't allow for strong conclusions regarding performance. Hybrid methods might be of interest for NIDS since they could allow to maintain both class level behaviour using logit matching and feature or relation loss to allow additional structure to be maintained, but the two hybrid studies are too small to generalise.

4.3 Loss Formulation and Performance Trade-Offs

As you can see in the analytical taxonomy there is no universally preferable distillation loss. Response-based KD has a low implementation cost and can be used by various families of teachers, but it suffers from the fact that the teacher's rationale is condensed into the output of the class. Feature matching transfers richer representations, but can take over training design, transfer nuisance features which are not useful to student. A relation matching refers to a looser structural target, which may work well for cross-architecture transfer, but may increase the cost of training, since these calculations are done at the batch level. Ablation studies isolating the effect of each term are needed if hybrid objectives are used to combine these benefits. In the future, the full objective, temperature T, supervised-distillation weight α, feature weights, teacher aggregation weights and whether they were tuned using validation data should be reported.

4.4 Hardware Evidence and the Meaning of "Lightweight"

The deployment evidence is the weakest side in corpus. Only 31% of the studies test physical edge hardware, and ranges of 20-500 KB, 0.07-2.1 ms, are not sufficient to make a meaningful comparison, without the context of the hardware. Raspberry Pi and NVIDIA Jetson boards have some differences in their processor architecture, the type of acceleration they support, the memory bandwidth and power profile. The performance of a model also varies with the numerical precision; an INT8 student may require much less memory and may perform differently from an FP32 model with the same architecture. To that effect, an exhaustive hardware table checking the parameters, model size, board, processor, precision, inference framework, batch size, latency, throughput, peak memory, and peak energy should be maintained. The studies where they are only working with a desktop latency report should describe their findings as a computational profiling and not a validation for edge deployment.

4.5 As an Adversarial Robustness Deployment requirement

One of the problems is that the frequency of adversarial testing is low, which is especially problematic for NIDS, which will be used in a hostile environment. A small student may have teacher vulnerabilities or develop new ones after compression if it is compressed to be the same size as the teacher. Robustness claims should thus rely on direct attack evaluation, explicitly state perturbation budgets and attack iterations, and state robust accuracy (or attack success rate) in addition to clean accuracy. When possible, studies should be performed comparing distilled students, undistilled students, and adversarially trained baselines, under the same threat model.

5. REVISED CONCEPTUAL FRAMEWORK

Figure 2 distinct between the overall KD evidence base and the ensemble subset, and distillation design to deployment and security assessment. The pipeline starts with a network traffic and goes to teacher supervision and teacher knowledge-transfer formulation then creates the lightweight student, and tests it on a defined edge platform.

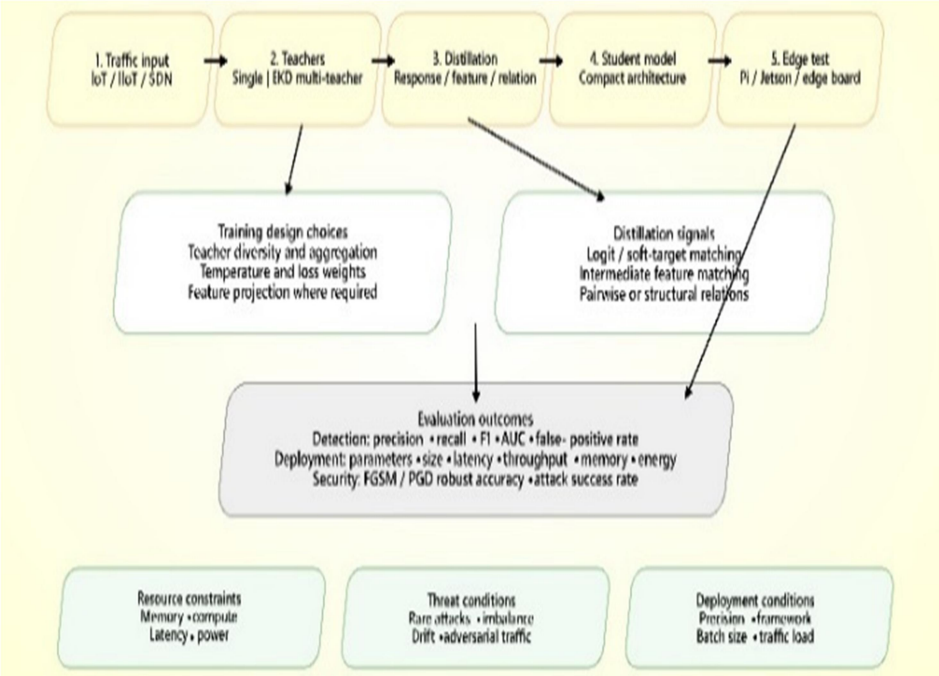

Figure 2. KD framework in lightweight NIDS, with EKD represented as the multi-teacher subset.

The teacher multiplicity and the type and aggregation of losses are regarded as teacher training design and the dimensions of evaluating are parameters, numerical precision, latency, memory, energy and robustness. Results may vary based on resource limits, threat conditions and deployment settings.

5.1 Research Schedule

First, single-teacher, homogeneous-ensemble and heterogeneous-ensemble supervision should be compared for the same data splits, student architecture and optimisation settings for future work. Secondly, research should go beyond disclosure of just the distillation method's name and there should be disclosure of the complete loss equation, that is, temperature, weighting, feature projections and teacher aggregation. Third, evaluation of hardware should be based on study level reporting templates, associated with a named edge platform and a numerical precision. Fourth, robustness testing is a desired part of NIDS evaluation and should be routinely incorporated in the process, particularly when NIDS are advertised for use in safety or hostile environments. Federated distillation and privacy-aware multi-teacher learning also should be explored as these approaches could enable distributed security models to share supervisory information without centralizing raw traffic.

6. CONCLUSION

A systematic review was performed on 42 studies of knowledge distillation for lightweight network intrusion detection. The evidence needs to be more widely distributed to include KD in its scope only 1 teacher in 30 studies, only 12 studies feature ensemble supervision. EKD should thus be called a 'working' paradigm and not the architecture of the whole corpus.

The most prevalent approach is response-based distillation, accounting for 76% of 42 studies, in part due to its simplicity and the fact that it is architecture agnostic and has a visible "transfer target" in the form of logits. Feature, relation and hybrid methods provide more powerful forms of supervision, but necessitates further tuning, projection or alignment. Only half of the studies evaluate physical edge robustness (31%) and adversarial robustness (21%). The most robust future research will weave between loss design, teacher diversity, per class detection behaviour and physical deployments conditions, in clearly stated hardware and security conditions. In order to reliably compare lightweight KD-based NIDS across studies, or judge them ready for edge deployment, such reporting is necessary.

6.1 Limitations of the Review

There are several limitations to this review. First, the synthesis is hinging on the reporting in the included studies and reporting across several variables was variable. Second, our search strategy could seem to be a bit broad, so there may be some other text that is relevant to distillation concepts but is not written in terms of distillation or ensemble learning. Third, quantitative meta-analysis was not performed due to significant variation in the outcome measures. Last, the review protocol was not registered publicly, which slightly detracts from the transparency of the process, although PRISMA 2020 reporting principles were adhered to strictly.

REFERENCES

1. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (pp. 108–116). https://doi.org/10.5220/0006639801080116
2. AlNomasy, N., Alamri, A., Aljuhani, A., & Kumar, P. (2025). Transformer-based knowledge distillation for explainable intrusion detection system. Computers & Security, 141, 104012. https://doi.org/10.1016/j.cose.2025.104012
3. Wisanwanichthan, T., & Thammawichai, M. (2025). A lightweight intrusion detection system for IoT and UAV using deep neural networks with knowledge distillation. Computers, 14(7), 291. https://doi.org/10.3390/computers14070291
4. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv:1503.02531. https://arxiv.org/abs/1503.02531
5. Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015). FitNets: Hints for thin deep nets. International Conference on Learning Representations. https://arxiv.org/abs/1412.6550
6. Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3967–3976). https://arxiv.org/abs/1904.05068
7. Quyen, N. H., Duy, P. T., Nguyen, N. T., Khoa, N. H., & Pham, V.-H. (2025). FedKD-IDS: A robust intrusion detection system using knowledge distillation-based semi-supervised federated learning and anti-poisoning attack mechanism. Information Fusion, 117, 102807. https://doi.org/10.1016/j.inffus.2024.102807
8. Xie, B., Wang, Z., Zeng, Z., He, D., & Chan, S. (2025). DTKD-IDS: A dual-teacher knowledge distillation intrusion detection model for the industrial Internet of Things. Ad Hoc Networks, 174, 103869. https://doi.org/10.1016/j.adhoc.2025.103869

9. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In 2016 IEEE Symposium on Security and Privacy (pp. 582–597). IEEE. https://doi.org/10.1109/SP.2016.41
10. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (pp. 39–57). IEEE. https://doi.org/10.1109/SP.2017.49
11. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations. https://arxiv.org/abs/1412.6572
12. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations. https://arxiv.org/abs/1706.06083
13. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71