International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/



(REVIEW ARTICLE)

Check for updates

ROBUSTNESS EVALUATION OF QUANTUM MACHINE LEARNING FOR NETWORK SECURITY: A SYSTEMATIC MAP AND CRITICAL APPRAISAL

Sarah Ali Abdullah *, Azzah Hazem Zaki and Saad Ahmed Dheyab

College of Engineering, University of Information Technology and Communications, Baghdad, Iraq.
* Corresponding Author

ORCID Details
Sarah Ali Abdullah: https://orcid.org/0000-0001-8695-7804
Azzah Hazem Zaki: https://orcid.org/0000-0001-6585-4531
Saad Ahmed Dheyab: https://orcid.org/0000-0003-0702-6679

International Journal of Science and Research Archive, 2026, 21(02), 267–293

Publication history: Received on 30 September 2026; revised on 09 November 2026; accepted on 11 November 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.21.2.1995

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

There is not much information about detecting robustness of quantum machine learning detectors even though the idea of using the latter for network attack detection has already been made. Articles from Scopus and IEEE Xplore databases were retrieved in English for research published after 2017 and searched in the last week of September of 2026. A two-level screening plan was applied with the first level mapping 363 abstract records and the second level performing a check of robustness claims made in abstracts against their corresponding full text articles and a final level assessment against a 14-item rubric. Overall, 204 reports were retrieved of which only 76 were selected for the analysis, and finally 33 studies were selected of which three were identified via other methods. Robustness was evaluated mainly as simulated quantum noise (17 studies) and as distribution shift within single benchmarks (10). Attacks on the detector were tested in nine studies; poisoning in one; and attacks such as backdoors or certified robustness were not found at all. Gradient masking was only partly implemented without an in-depth check; also, not a single attack achieved the level of being realizable and optimal for traffic simultaneously, and not a single noise or hardware study was able to perturbed a classical baseline in the same way in that exact way. Significance testing was absent from 25 studies, and 25 relied on simulation alone. When an abstract claimed to be robust without naming a test, full texts of 18 out of 24 articles were found not to have any test reported, and none of the articles had their conclusions strictly supported by the evidence when judged under an explicit adjudication rule. The method used is not free from errors since it is based on AI-assisted screening, driven retrieval mainly through access, and also appraiser's poor agreement on the claims item. The argument of QML that makes network detectors more robust is not supported at the moment; a minimum evaluation and reporting checklist as well as a research roadmap are proposed to help.

Keywords: Adversarial Machine Learning, Intrusion Detection, Network Security, Quantum Computing, Quantum Machine Learning, Quantum Noise, Robustness, Systematic Review.

1 INTRODUCTION

Machine-learning models have become the detection engine of anomaly-based network intrusion detection systems (NIDS) [1], yet they were shown to be sensitive to perturbations crafted at training or test time [2]. Against Kitsune, gray- and black-box attacks mounted in traffic space reached evasion rates above 97% in half of the tested cases [1]; threat models that grant the attacker full knowledge of the detector were judged unrealistic [3]; and evaluation pitfalls were found to be widespread in 30 papers from top-tier security venues [4]. A robustness result is only as informative as the evaluation behind it.

Quantum machine learning (QML) detectors have since been proposed for the same tasks, built from quantum kernels [5]–[7], variational and hybrid quantum–classical networks [5], [8], generative models [9] and federated designs [10], often with robustness as part of the motivation. Quantum adversarial machine learning was presented as a possible source of quantum advantage [11], and one intrusion-detection study assumed that quantum techniques resist adversarial examples better than classical ML [12]. The experimental record is less favourable. Quantum classifiers were vulnerable to crafted adversarial examples [13], were more vulnerable than classical models in white-box settings [14], and resisted label flipping but not gradient-based evasion [15]. The role of device noise is disputed: a robustness bound that improves with depolarising noise was derived [16], whereas in experiments noise was an unreliable passive defence [15] and a moderate noise level sometimes maximised attack success [17]. Most of this evidence came from image benchmarks [13], [15]; whether it holds for network detectors depends on how they were evaluated. QML has also been used as an attacker: a quantum generative adversarial network evaded a Random Forest NIDS in 95.3% of cases [18].

Earlier reviews of QML for intrusion detection were organised around algorithms, datasets and metrics [19]–[22] or around threat intelligence [23], while robustness syntheses covered QML in general [15], [24] (Table 1). They agreed that simulation dominates, at 80.6% [22] and 83.3% [23] of the studies reviewed, and later reviews called for standardised protocols [19] and hardware-validated robustness evaluation [24]. Study quality was appraised only with a general-purpose checklist [22], and the closest precedent for auditing claims against evidence comes from quantum learning in power systems [25]. How robustness has been evaluated for QML detectors of network attacks, and how rigorously, has not been examined.

Table 1: Earlier Reviews Compared With the Present Review

Review	Design and corpus	Domain	Stated focus
[11] M. T. West et al. (2023)	Perspective	Quantum adversarial ML (general)	Recent progress, key challenges and research directions; robustness as a possible source of quantum advantage
[21] O. K. Nicesio et al. (2023)	Systematic literature review (articles 2017–2022)	QML for network intrusion detection	Algorithms proposed; benefits reported over classical models
[20] O. Faker and N. E. Cagiltay (2024)	Systematic mapping review (21 of 240 studies, four databases)	QML for intrusion detection	Increasing the efficiency of ML-based anomaly detection
[26] V.-L. Nguyen et al. (2025)	Magazine article on adversarial threats	QML-assisted 6G network functions and digital twins	Kernel poisoning, backdoors and adversarial noise; candidate defences
[19] A. Kaissar et al. (2026)	Systematic survey (peer-reviewed studies up to January 2026)	QML for network intrusion detection, including quantum-inspired models	Algorithms, datasets and evaluation metrics
[22] N. Louise and V. Kavitha (2026)	PRISMA 2020 systematic review (87 studies, 2018–2024; CASP quality appraisal)	Hybrid quantum–classical ML for intrusion detection, with emphasis on the Store Now, Decrypt Later threat	Accuracy versus deep learning; simulation reliance, latency and trainability barriers
[23] P. Rezaei et al. (2026)	PRISMA-ScR scoping review (24 works, 2020–2025)	QML for cyber-threat intelligence	QML frameworks across four threat-intelligence domains; simulator reliance

[24] Y. A. Kustiawan and K. I. Ghauth (2026)	PRISMA-ScR scoping review (53 empirical studies, 2020–2026)	Adversarial robustness in QML in general	Threat models, defences, evaluation approaches and practical constraints
[15] S. R. Nowmi et al. (2026)	Systematization of knowledge with experiments (five attacks, three threat models)	Adversarial robustness in QML; experiments on MNIST and AZ-Class	Accuracy–robustness trade-off; noise as a passive defence
[25] L. Tightiz and E. Oh (2026)	Critical appraisal (59 records: 57 application studies, 2 reviews)	Quantum learning in power systems	What studies claim versus what they demonstrate; study-level audit
This review	PRISMA 2020 systematic map (363 records) with critical appraisal of 33 studies (Scopus, IEEE Xplore)	QML detectors for network security	How robustness was evaluated, across all robustness dimensions; 14-item rigour rubric; claims checked against evidence at abstract and full-text level

Cells summaries each review's abstract; "Stated focus" is what the abstract says the review analysed.

The review covers models whose inference passes through a parameterised quantum circuit or a quantum kernel, including hybrid models, applied to the detection of attacks in network traffic or logs, including attacks on quantum key distribution (QKD) links. A robustness evaluation is any test beyond clean, in-distribution data: adversarial, noise-related, distributional or privacy-related (Table 2). Four research questions (RQs) were posed. RQ1: Which robustness dimensions were evaluated, and in what settings? RQ2: How rigorous were these evaluations against published guidance for adversarial and security evaluation? RQ3: Do robustness claims hold up, both when abstracts are compared with full texts and when conclusions are compared with reported results? RQ4: Which minimum practices and research directions follow?

Following PRISMA 2020 [27], 363 records were mapped at abstract level and 33 studies were appraised in full. The contributions are a systematic map and taxonomy of robustness evaluations (Sections IV, V); a 14-item rigour rubric grounded in guidance on adversarial evaluation [28], realistic NIDS attacks [3], [29] and security pitfalls [4], with every extracted value backed by a verbatim quote (Sections III, VI); a two-level audit of robustness claims (Section VI); and a minimum evaluation and reporting checklist with a research roadmap (Sections VIII, IX).

2 BACKGROUND

2.1 Models, Encoding and Execution

Variational classifiers and quantum kernel estimators both use the quantum state space as the feature space [30]. A feature vector $x \in \mathbb{R}^d$ is mapped by an encoding circuit $U(x)$ to an n-qubit state,

$$|\varphi(x)\rangle = U(x)\,|0\rangle^{\otimes n}, \tag{1}$$

after which a variational classifier applies a trainable circuit and measures an observable, while a kernel method computes state overlaps on the device and passes them to a classical support vector machine [30]. Angle encoding assigns one feature per qubit and was used in a 6-qubit obfuscation layer [31]; amplitude encoding writes a normalised vector into the amplitudes,

$$|\varphi_{amp}(x)\rangle = \frac{1}{\|x\|_2}\sum_{i=0}^{2^n-1} x_i\,|i\rangle, \tag{2}$$

and was used to embed 16 principal components of NSL-KDD on four qubits [32]. Because (2) divides by $\|x\|_2$, inputs that differ only in scale map to the same state, so the magnitude of a traffic-feature vector is lost before classification.

Noise was expected to limit the circuits that noisy intermediate-scale quantum (NISQ) devices can execute reliably [33]. Results therefore depend on the execution setting: statevector simulation, shot-based simulation, simulation with device noise models, or hardware. Shot-based outputs are stochastic, which matters for gradient-based attacks (Section VI-A). Idealised noise is usually modelled with channels such as the depolarising channel,

$$\mathcal{D}_p(\rho) = (1-p)\rho + p\,\frac{I}{2^n}, \tag{3}$$

applied, for example, with probabilities up to 0.3 in [17]. Training adds a further limit, since gradients of many parameterised circuits vanish exponentially with qubit count [34], also under local noise [35]. The included detectors used four to sixteen qubits [7], [9], [31], [32], which keeps simulation tractable but limits extrapolation to larger devices.

2.2 Fair Comparison and Robustness

Whether beating classical machine learning should remain the dominant narrative of QML has been questioned [36]. Classical models outperformed quantum classifiers across 160 datasets [37], were competitive even on problems tailored to quantum models [38], and act as surrogates for large classes of re-uploading models [39]. On four NIDS datasets, tuned Random Forest and XGBoost matched or exceeded quantum models, a parity traced to classical preprocessing and regularisation [40], and none of 57 power-system studies offered a matched-budget classical baseline [25]. Robustness has been proposed as a separate route to advantage [11], [41], but the same standard applies: a quantum detector is more robust only if a comparably tuned classical baseline degrades more under the same test (items B6 and B7, Section III-C).

Let $\hat{y}(x)$ be the label that a detector assigns to x. An adversarial example is an input $x' = x + \delta$ with $\|\delta\|_p \leq \varepsilon$ for which $\hat{y}(x') \neq \hat{y}(x)$, and the detector is ε-robust at x if no such perturbation exists,

$$\hat{y}(x+\delta) = \hat{y}(x), \quad \forall \delta: \quad \|\delta\|_p \leq \varepsilon. \tag{4}$$

Robustness can be estimated empirically by attacking the model, or certified by proving (4) for every perturbation within the budget; in QML the empirical route dominates [24]. A large state space does not protect a classifier, since the perturbation needed for misclassification shrinks as dimensionality grows [42], and both classifiers run on superconducting qubits [43] and quantum-kernel support vector machines [44] were deceived by small perturbations. Guarantees exist only under conditions, through quantum hypothesis testing [45], protection of in-distribution data [41] and Lipschitz bounds tied to the encoding norm [46]. Depolarising noise (3) yields a bound that improves with noise strength [16] and suitable noise channels are equivalent to randomised smoothing [47], but whether noise protects a deployed detector is an empirical question (Section VI-B). Network traffic adds constraints of its own, because attacks not designed for communication protocols may produce unrealistic examples [29].

3 REVIEW METHODOLOGY

The review is reported according to PRISMA 2020 [27] and PRISMA-S [48]; operational details are given in Supplementary Methods S1.

3.1 Protocol, Search and Eligibility

The review was not registered, and no protocol was published before the search. Criteria and screening codes were determined by means of the screening workbooks, in which every decision and change was recorded and dated. Scopus and IEEE Xplore were searched on 28 September 2026 (IEEE Xplore) and 29 September 2026 (Scopus) for publications from 2017 onward, combining QML terms with network-security terms (full strings in Supplementary Methods S1). Studies found through citation and arXiv searching were assessed with the same criteria. Table 2 gives the eligibility criteria; DGA botnet detectors were flagged as borderline, and QKD-link detectors formed a separate stratum. Two amendments were made. One was made on 19 October 2026 when after title and abstract screening a robustness analysis became an eligibility factor, while all previous codes were preserved. The other one was in November when the situation with some very valuable reports made them impossible to access and a two-step review scheme was put in place with a claims audit linking them (Section III-D) this post hoc decision made no changes to the criteria, codes, extraction data or earlier decisions. The screening worksheet and the judgment matrix that have been used during the work are available as supplementary information.

Table 2: Eligibility Criteria

Criterion	Included	Excluded
Model	QML model whose inference passes through a parameterised quantum circuit or a quantum kernel, on hardware or in simulation, including hybrid quantum–classical models	Quantum-inspired classical methods; models based on quantum annealing; quantum methods used only for optimisation; classical models; cryptographic protocols without learning
Task	Detection of attacks in network traffic or logs: intrusion and anomaly detection, DDoS and botnets (DNS/DGA flagged as borderline), threats in encrypted traffic, attacks on CPS, ICS, smart-grid,	Applications outside security; security functions other than detection (for example, cryptography, authentication, privacy protection); detection outside network security

	vehicular, IoT or 5G networks, and attacks on QKD links	(for example, malware files, phishing, spam, code vulnerabilities)
Evaluation	A robustness evaluation of the detector: adversarial evasion, poisoning or Byzantine behaviour, quantum or hardware noise, distribution shift or unseen attacks, privacy or model extraction; named or claimed in the title or abstract and present in the full text	No robustness evaluation; no detection result for the QML model
Study type	Primary study with an evaluated model	Overviews, position papers and frameworks without an evaluated model; reviews and surveys (used only as background); non-primary items such as editorials and conference reviews
Language and date	English; published from 2017 onward	Other languages; published before 2017

3.2 Selection

Titles and abstracts were screened by an AI tool (R1; Claude Opus 5.5, Anthropic) that recorded a justification for each decision. A second AI tool (R2; ChatGPT, OpenAI; version not recorded) screened a seeded random sample of 441 records (20% per database), and agreement was measured with Cohen's kappa [49] on the Landis–Koch scale [50]. A human reviewer, Tariq Adnan Fadil (Ministry of Higher Education and Scientific Research Iraq, Studies, Planning and Follow-up Directorate), blind to the R1 codes, screened 80 records excluded by both tools, resolved 37 disagreements and checked the robustness criterion for 24 records; his decisions were final. A stopping rule fixed in advance required re-screening of the remaining exclusions under a code if any check overturned an R1 exclusion. Full texts were sought until 2 November 2026 and assessed by the first author from an AI-assisted reading, with borderline decisions flagged.

3.3 Data Extraction and Critical Appraisal

Sixteen items were extracted per study (Supplementary Table S2). Extraction was AI-assisted, each value was backed by one to three verbatim quotes verified by script against the source, and the first author checked every extracted value against its quotes. The nine studies included on 2 November 2026 were extracted with a reduced field set. Results were synthesised as counts; no effect was pooled.

Rigour was appraised with the 14-item rubric in Table 3, which turns guidance on evaluating adversarial defences [28] and on pitfalls of learning-based security [4] into checks. Items B1–B4 target known failures such as obfuscated gradients [51], defences broken by adaptive attacks [52] and poorly tuned attacks [53]; B5 checks network constraints [54]; and the remaining items cover baselines, replication, statistical testing [55], dataset flaws [56]– [58], metrics, hardware, code and whether conclusions stay within the evidence (B14). Items were coded Yes, Partial, No or N/A, with unreported items coded No. The human reviewer independently appraised 18 studies included at that stage, and the 36 disagreements were resolved by consensus, with single-appraisal codes checked by the first author; the 15 studies included later carry a single appraisal of B1–B13. Because B14 is a judgement, it was appraised a second time, blind, for the remaining 22 studies, and disagreements were adjudicated with an explicit rule: No when a conclusion is contradicted by the study's own results or concerns an untested condition; Partial when it is supported but generalised beyond what was tested, or credits an untested mechanism; Yes, when it is confined to what was tested. The rule changed none of the 11 earlier double-appraised codes. The adjudicated codes are the primary result, with the first-appraisal codes as a sensitivity analysis.

Table 3: Rigour Rubric Used for the Critical Appraisal

Item	Criterion	Grounding
B1	Threat model stated explicitly (attacker knowledge, goal, capabilities)	[3], [28], [59]
B2	Attack stronger than single-step FGSM (iterative, adaptive, ensemble or transfer attacks)	[52], [53], [60], [61]
B3	Attack budget reported and varied (sweep of ε or of attack strength)	[28]
B4	Checks against gradient masking (adaptive or expectation-over-transformation attacks, sanity checks)	[51], [53], [62]

B5	Perturbations respect network-traffic constraints (problem space or constrained features)	[29], [54], [63]–[65]
B6	Classical baseline evaluated under the same robustness test	[37], [39]
B7	Baseline tuning or matched model budget reported	[25], [37]
B8	Multiple runs or seeds, with variance reported	Basic experimental practice
B9	Statistical significance testing	[55]
B10	Dataset issues acknowledged (age, known flaws, imbalance), or a current dataset used	[56]–[58]
B11	Metrics beyond accuracy reported under attack or noise (F1, recall, false-positive rate, AUC)	[24], [66]
B12	Real quantum hardware used, and at what scale	[33], [67]
B13	Code or data released (link given)	Reporting item
B14	Conclusions stay within the evidence (reviewer judgement; the paper's conclusion is quoted; adjudicated by the rule in Section III-C)	[28], [54]

Codes: Yes, Partial or No from the full text, where Partial means that a criterion is met for only part of the evaluation or is reported only in part; an item a paper does not report is coded No. B1–B5 are N/A when a study runs no attack (evasion, poisoning, extraction or inversion); B6 and B11 are N/A only when a study has no robustness test.

3.4 Two-Tier Design and Claims Audit

The first-tier maps 363 records: the 204 reports sought at full text and the 159 exclusions for lack of a robustness evaluation that carried a human answer. Model families and datasets were matched automatically in titles, abstracts and keywords, so they record what abstracts mention. Each abstract received a robustness signal (Named, when a test was described; Claimed only, when robustness was asserted without a test; or Neither), taken from the human reviewer for 290 records and from R1 for 73. The second tier is the appraisal of the included studies. A claims audit links the tiers at two levels: the abstract signal of each assessed report is cross-tabulated with its full-text outcome, and B14 records whether each study's conclusions stay within its evidence. Because full texts were obtained by access, and more often for abstracts that named a test, audit proportions hold for the assessed reports only.

4 RESULTS I: STUDY SELECTION AND SYSTEMATIC MAP

4.1 Study Selection

The searches returned 2,978 records (Fig. 1). After removal of 554 duplicates and 216 ineligible items, 2,208 records were screened, and 1,847 were excluded, 384 by a human decision and 1,463 by the AI tools alone, reported separately as PRISMA 2020 requires when automation tools are used [27]. On the double-screened sample, the tools agreed on 91.6% of records ($\kappa = 0.598$, moderate [50]). The human checks overturned part of the AI decisions, including 12 R1 exclusions among the 37 disagreements, so the stopping rule was applied. Re-screening returned to full-text assessment 9 of 167 records excluded for task reasons (Part A) and 104 of 243 records excluded for lack of a robustness evaluation in the abstract (Part B). A planned full-text validation of 30 robustness exclusions was stopped after 3 records, none of which reported a robustness evaluation.

Of 204 reports sought, 128 could not be retrieved by the closing date. Of the 76 assessed, 46 were excluded (Supplementary Table S4), most because they reported no robustness evaluation of the QML detector (32), even when their authors called the model robust [68]. Thirty were included and three more were found through other methods [6], [40], [69], giving 33 studies. Seven were borderline inclusions, removed in the sensitivity analysis: two whose only evidence was a hardware–simulation comparison [7], [70], one with partial noise testing [71], one with an undescribed zero-day protocol [72], and three at the edge of the task scope [73]–[75].

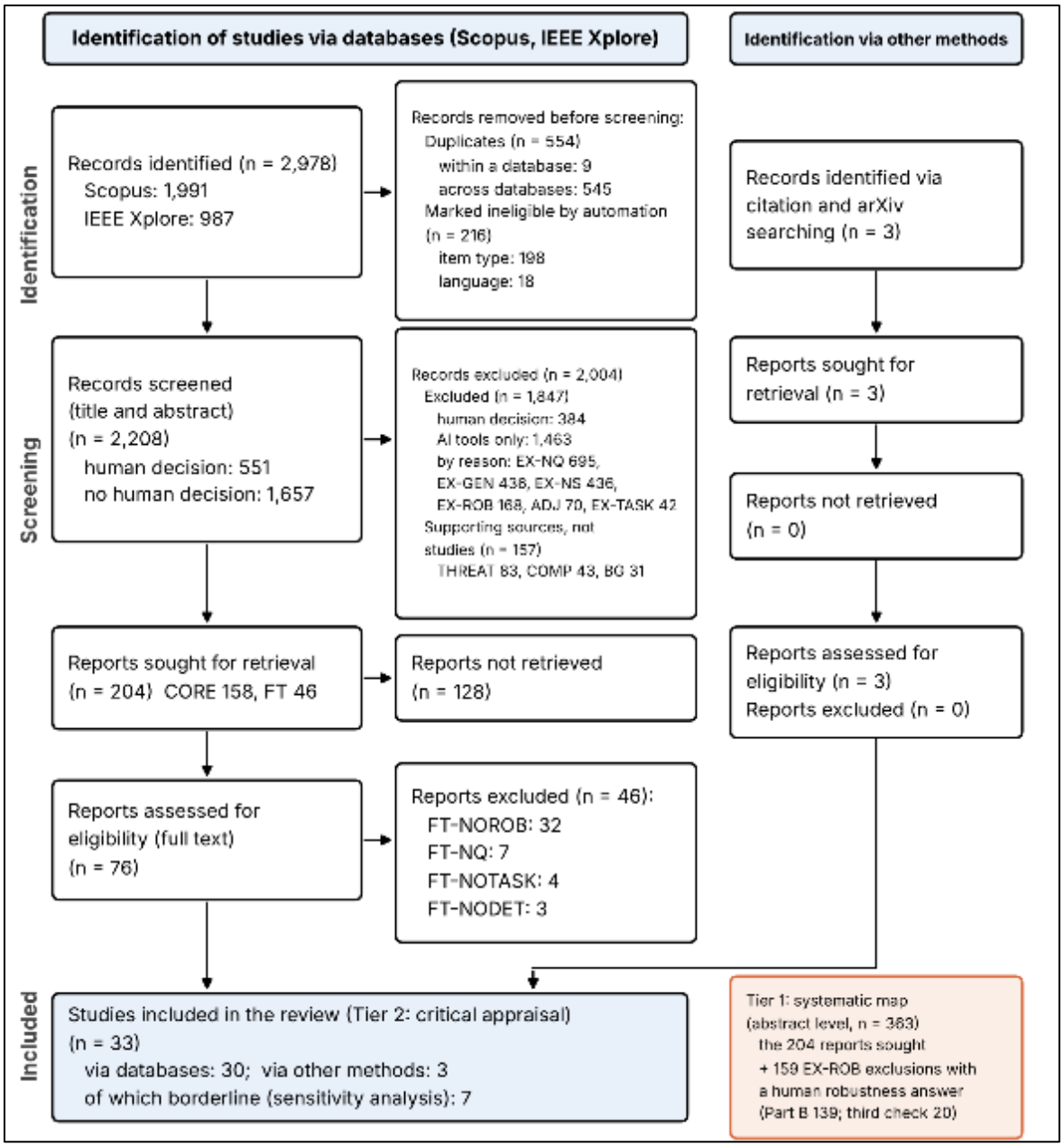


Figure 1: PRISMA 2020 flow diagram of the review Human and AI decisions are only separated at screening, just *when* it is mandated if automation tools are to be used [27]. Supporting documents (THREAT, COMP, BG) were cited but not combined. The Tier 1 map fuses the 204 reports retrieved with the 159 EX-ROB exclusions that carried a human robustness answer (139 from Part B of the stopping rule and the remaining 20 from the third human check); Tier 2 comprises the 33 final studies.

4.2 Systematic Map (Tier 1)

The field is young: 274 of the 363 records (75%) were published in 2025 or 2026 (Fig. 2a). Variational circuits and quantum neural networks were the most frequently named family (168 records), followed by quantum kernels (72) (Fig. 2b), and 220 abstracts (61%) named no common benchmark dataset. Among the 204 full-text candidates, 76 abstracts named a robustness test and 105 only claimed robustness, 94 of the latter from 2025 or 2026. Where a type was named, simulated quantum noise was the most frequent (27 mentions), followed by hardware (15), distribution shift (14), adversarial evasion (9) and poisoning (4) (Fig. 2c). Retrieval depended on the signal: 59% of Named records were assessed in full, against 23% of Claimed-only records (Fig. 2d).

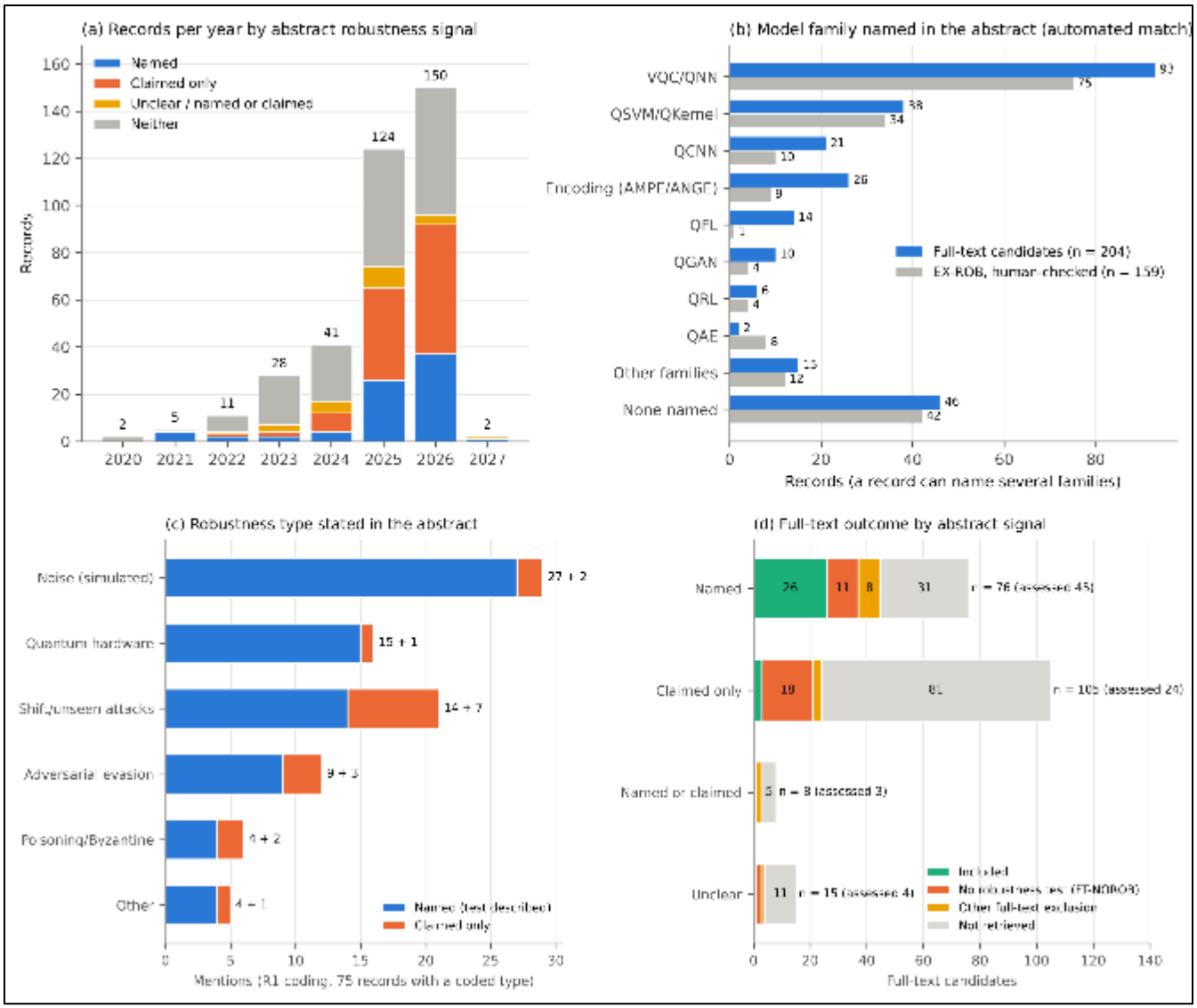


Figure 2: Systematic map of the Tier 1 records (n = 363). (a) Records per publication year by abstract robustness signal (2027 = records dated 2027 at export). (b) Model family named in the abstract, by stratum (automated match; a record can name several families). (c) Robustness type stated in the abstract, as coded by R1 for 75 records (blue: a test is described; orange: robustness is only claimed). (d) Full-text outcome of the 204 full-text candidates by abstract signal. Blue counts follow R1's type coding and include two records whose human-coded signal was Unclear, so they exceed the Named counts of Fig. 4 by one for noise, hardware and adversarial evasion. Source: screening workbook, sheets Map_Tier1 and Claims Audit.

4.3 Included Studies (Tier 2)

The 33 included studies (Table 4) were recent, 20 of them from 2026. Most addressed intrusion or anomaly detection on flow or packet features; others targeted power and cyber-physical systems, 5G, telephony denial of service and DGA botnets. NSL-KDD [56] was the most used dataset (8 studies), followed by UNSW-NB15 [76] (6) and CIC-IDS2017 [77] (5), with TON_IoT [78] among those used in four. Execution was overwhelmingly simulated: 25 studies ran nothing on quantum hardware.

Table 4: Characteristics of the Included Studies (n = 33)

Ref.	Study	Task; data	QML model	Execution	Robustness evaluated
[5]	Cirillo et al. (2026)	Intrusion detection; ToN-IoT, NSL-KDD	Pegasos-QSVC, VQC, HQNN	Ideal and noisy simulation (three IBM fake backends)	Hardware noise (simulated)
[6]	Cotrupi and Callahan (2025) †	Binary NIDS; BCCC-CSE-CIC-IDS2018 (n = 60)	Quantum-kernel SVM	IBM 127-qubit device; noiseless simulator	Run-to-run variation over ten hardware runs
[7]	Badami (2026) *	ICS anomaly detection; SWaT, HAI	QSVM, ZZ feature map, 8 qubits	Statevector simulation; ibm_fez for kernel elements only	One hardware-versus-simulation kernel comparison
[8]	Wang et al. (2025)	Malicious-traffic detection; 5G-NIDD	Five hybrid QCNNs	Simulation (CPU)	Cross-base-station shift; leave-one-attack-type-out
[9]	Malwi et al. (2026)	Zero-day detection; CICIDS2017	QCBM, 16 qubits	Noiseless simulation	Single-family zero-day hold-out
[10]	Vyas et al. (2026)	Binary IIoT NIDS; WUSTL-IIOT-2021, ToN-IoT, SCVIC-APT-2021	Hybrid Q-CNN trained by QFL (10 clients)	Simulation, ideal and depolarizing	Depolarizing noise (simulated)
[31]	Rawas et al. (2026)	Binary NIDS; UNSW-NB15, NSL-KDD	PQC obfuscation layer with classical classifiers, 6 qubits	Ideal statevector simulation	Model extraction; feature inversion; Gaussian input noise
[79]	Kukliansky et al. (2024)	Binary NIDS; NF-UNSW-NB15	QNN, six XY layers	IonQ noise-model simulation; IonQ Harmony and Aria-1	Hardware noise (noise models and QPU runs)
[12]	Qi et al. (2024)	Binary NIDS; KDD Cup 99, UNSW-NB15 subsets	VQCNN, 10 qubits	Noiseless simulation	FGSM evasion
[40]	Gillani et al. (2026) †	Binary NIDS; NSL-KDD, UNSW-NB15, CICIDS2017, NF-ToN-IoT-v2	Six hybrid VQCs, two QSVMs	Simulation, incl. density-matrix noise	Simulated noise channels; KDDTest+ shift
[32]	Thomos and Andronikos (2026)	Binary and four-class NIDS; NSL-KDD	VQC, 4 qubits	Ideal statevector simulation	Unseen attacks (KDDTest+)
[80]	Hammadia et al. (2026)	FDIA versus fault in substation relays; simulated HYPERSIM data	Variational circuits, 3–6 qubits	Noiseless simulation; hardware-in-the-loop testbed	Measurement noise; CT errors; time-shift attack
[69]	Filardo (2026) †	Binary NIDS; NSL-KDD, UNSW-NB15 subsets	Stochastic QNN, 4 qubits	Exact density-matrix simulation	White-box FGSM and PGD; depolarizing and dephasing noise

[70]	Alomari and Kumar (2023) *	Binary LDAP-DDoS detection; CIC-DDoS2019	Pegasos QSVC, ZZ feature map	Statevector simulation; ibm_oslo	One hardware run versus simulator
[71]	Danasabe et al. (2026) *	Multiclass SDN IDS; CSE-CIC-IDS2018, OD-IDS2022	Hybrid VQC and TabNet (QATNet)	Simulation; fake-backend noise on embedded features	Simulated noise on the embedding stage only
[72]	Shaw et al. (2026) *	NIDS and response; NSL-KDD, CIC-IDS2017, CSE-CIC-IDS2018	Hybrid QML with quantum RL (NE)	Simulation, 4–8 qubits	Zero-day attacks (protocol not described)
[73]	Suryotrisongko et al. (2022) *	Botnet DGA detection; authors' DGA dataset	Hybrid quantum-classical network	Simulation with four IBM noise models	FGSM, PGD, BIM; device noise (simulated)
[74]	Suryotrisongko and Musashi (2022) *	Botnet DGA detection; authors' DGA dataset	Hybrid quantum-classical network (NE)	Simulation with eight IBM noise models	Device noise (simulated), n = 100
[75]	Peng et al. (2024) *	Fuzzed-layer detection in 5G protocol logs; own srsRAN logs	Hybrid QNN (NE)	Simulation; IBM Kyoto	Four noise channels; hardware versus simulation
[81]	Ngo et al. (2026)	Six-class smart-grid anomaly detection; ICS power-system dataset	Partitioned QNN (NE)	Simulation, four 5-qubit circuits	FGSM and PGD (bounds 0.05, 0.1)
[82]	Hammami et al. (2025)	DDoS anomaly detection in traffic time series; CSE-CIC-IDS2018	QGAN built from VQCs (NE)	Simulation; Fake Vigo noise	Simulated noise (one fake backend)
[83]	Addo et al. (2025)	Smart-grid attack detection; synthetic multi-agent data, IEEE 39-bus case	Federated VQC, 8 qubits	Statevector and noisy simulation	Bounded input perturbation; Byzantine agents; simulated noise
[84]	Ngobeni and Kabaso (2026)	5G roaming-signalling anomalies; own testbed traces	Hybrid classical-quantum classifiers (NE)	Simulation	MITM feature-space evasion
[85]	Rosa-Remedios et al. (2026)	TDoS early warning; PSAP call records and synthetic scenarios	Fidelity quantum kernel (NE)	Simulation	Unseen and zero-day-like TDoS variants
[86]	Tscharke et al. (2026)	Semi-supervised anomaly detection; 11 datasets incl. KDD, DoH	QSVR, fidelity kernel, 5 qubits	IBM 27-qubit device (clean benchmark only); simulation	PGD evasion; simulated noise channels; adversarial training
[87]	Venkateswarlu et al. (2026)	Binary Edge-IoT anomaly detection; CICIoT2023	Federated hybrid VQC, 4 qubits	Simulation with depolarizing noise	Depolarizing noise with and without ZNE
[88]	Zhao et al. (2026)	Five-class traffic classification; NSL-KDD	Distributed QML framework (NE)	Simulation with sampled noise	Depolarizing noise and link failures (0× to 10×)
[89]	Verma and Bhatt (2026)	Cloud IoT/IIoT NIDS; CIC-IoT 2022/2023, Edge-IIoTset	Quantum self-attention encoder (QSAFE-Net)	Simulation; four-qubit subset on IBM Heron R2	Cross-dataset transfer (UDA, no target access)

[90]	Abreu et al. (2025)	Binary and multiclass IDS; UNSW-NB15, CIC-IDS17, CICIoT23, ToN-IoT	Several QML models (NE)	Seven IBM backends (device runs or noise models not stated)	Hardware noise across backends and transpilation levels
[91]	Spadari et al. (2025)	Misuse detection, attack classification; IoT-NIDD, Edge-IIoT	Two QNNs (amplitude, angle embedding)	Ideal simulation	Cross-dataset evaluation
[92]	Agarwal et al. (2026)	SDN-IoT IDS; NSL-KDD and three non-network sets	VQC, 9 qubits, in a SHA-256 and decision-tree pipeline	MATLAB simulation	One NSL-KDD attack class held out
[93]	Saxena et al. (2026)	Three-class DDoS detection; APA-DDoS	Overlap-based classifier and QSVM-QNN	Simulation, noiseless and six noise channels	Simulated noise channels
[94]	Sabeshuly et al. (2025)	Six-class traffic classification; CIC-IDS2017 and CIC-DDoS2019	Hybrid VQC, one-vs-rest	Simulation, noiseless and depolarizing	Depolarizing noise (simulated)

* Borderline study, removed in the sensitivity analysis. † Identified via other methods. NE: not extracted (reduced extraction of 2 November 2026). Values summarise the evaluation matrix (Supplementary File), where each value is backed by verbatim quotes.

5 RESULTS II: TAXONOMY OF ROBUSTNESS EVALUATION

Simulated quantum noise was the most frequently evaluated dimension (17 studies), followed by distribution shift or unseen attacks (10), adversarial manipulation (8) and hardware execution (6, plus 2 partial); poisoning and privacy attacks appeared once each [31], [83] (Table 5). Twenty-one studies examined a single dimension. The same ordering appeared in the abstracts (Section IV-B).

5.1 Attacks on the Detector

Only nine studies (27%) attacked the model [12], [31], [69], [73], [80], [81], [83], [84], [86]. Six ran gradient-based evasion, mostly FGSM, with PGD in four [69], [73], [81], [86]. Two manipulated inputs without gradients, through a time shift with false-data injection [80] and a man-in-the-middle in transit [84], and one tested model extraction and feature inversion [31]. Training-time attacks were nearly absent: one federated study placed two Byzantine agents against trimmed-mean aggregation [83], and no study tested a backdoor, although such threats have been mapped for QML [15] and QML-assisted network functions [26]. Budgets were varied in six of the nine studies. The perturbation domain separated the studies sharply. The six gradient-based attacks perturbed extracted features, in two cases four- or five-dimensional PCA inputs [69], [86] that the network never carries, while the two that manipulated traffic or signals [80], [84] used fixed manipulations that were not optimised against the detector. No study therefore tested an attack that was both strong and realisable, although network constraints affect attack success significantly [63] and can invalidate up to 80.3% of adversarial examples [54].

5.2 Quantum Noise and Hardware

Of the 17 noise studies, ten applied parametric channels and eight of those swept the noise strength; seven used noise models of named devices and compared fixed profiles, leaving the relation between noise level and detection performance unmeasured. Some details were missing, such as the probability of a single depolarising setting [87], [94] or whether noise was present during training [10], [93]. Hardware runs were small: a 27-qubit device ran only the clean benchmark, on 30 training and 50 test samples per dataset [86], only a two-layer model fitted on IonQ Aria-1 [79], and other studies ran ten circuits, single runs or small circuit subsets [6], [7], [70], [75], [89]. In three studies the hardware run was the only robustness evidence [6], [7], [70]. This reliance on simulation matches earlier reviews [19], [23] and the observation that robustness is usually evaluated under fixed execution conditions [67].

5.3 Shift, Certification and Federated Settings

The ten shift studies used at least five protocols: attack types held out within one dataset [8], [9], [72], [92], the official NSL-KDD test split [32], [40], synthetic attack variants [85], transfer between locations or corpora [8], [89], [91] and a change of execution environment [84]. Only two moved across datasets [89], [91], and test-set composition varied: the unseen-attack sets of [8] contained no benign traffic. Results reported as "robustness" thus came from tests of different difficulty. No study reported a certified bound or formal verification, and the available tools [16], [45]– [47] were not

applied. Of four federated or distributed studies [10], [83], [87], [88], three tested noises only, and none ran a privacy attack, although secure quantum federated learning remains an open requirement for deployment [26].

6 RESULTS III: CRITICAL APPRAISAL

Table 5 and Fig. 3 give the rubric codes. The median study met two applicable items in full (range 0–7), and the two highest scorers were shift studies that released their code [32], [40]. Agreement between appraisers on the 14 double-appraised studies was 78.3% (κ = 0.71 [49]). On B14 it was poor before adjudication: 7 of 22 studies (κ = −0.06), with the second appraiser coding four studies Yes [75], [82], [89], [90].

Table 5: Rubric Codes (B1–B14) and Robustness Dimensions of the 33 Included Studies

Ref.	Study	Dim.	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	Yes
Attack on the model																	
[86]	K. Tschark e et al. (2026)	AE, SN, HW	P	Y	Y	N	N	N	N	N	N	Y	Y	Y	N	P	5
[80]*	T. Hamma dia et al. (2026)	AE, IN	P	N	Y	N	Y	N	Y	Y	N	P	Y	N	N	P	5
[69]	G.-E. Filardo (2026)	AE, SN	P	Y	P	P	N	P	N	Y	Y	P	N	N	Y	P	4
[84]*	T. Ngobeni and B. Kabaso (2026)	AE, DS	P	N	N	N	Y	Y	N	P	P	Y	Y	N	P	P	4
[31]	S. Rawas et al. (2026)	PX, IN	Y	N	P	N	N	Y	P	P	P	P	Y	N	N	N	3
[12]*	H. Qi et al. (2024)	AE	N	N	Y	N	N	Y	N	N	N	N	Y	N	N	N	3
[73]†	H. Suryotri songko et al. (2022)	AE, SN	N	Y	Y	N	N	N	N	P	Y	P	N	N	P	N	3
[81]*	H.M. Ngo et al. (2026)	AE	P	Y	Y	N	N	Y	P	N	N	P	N	N	N	P	3
[83]*	K. Addo et al. (2025)	AE, PB, SN, FD	P	N	Y	N	N	P	N	P	N	P	P	N	N	N	1
Distribution shift or unseen attacks																	
[32]	A. Thomos and T.	DS	–	–	–	–	–	Y	Y	Y	Y	Y	Y	N	Y	P	7

	Andronikos (2026)																
[40]	S. A. Gillani et al. (2026)	SN, DS	–	–	–	–	–	P	Y	Y	Y	Y	Y	N	Y	P	6
[89]*	R. Verma and R. Bhatt (2026)	DS, HW (p)	–	–	–	–	–	P	P	Y	N	Y	Y	P	Y	P	4
[9]	W. A. Malwi et al. (2026)	DS	–	–	–	–	–	Y	Y	N	P	P	Y	N	P	N	3
[85]*	C. Rosa-Remedios et al. (2026)	DS	–	–	–	–	–	Y	N	P	N	Y	Y	N	N	P	3
[8]	Z. Wang et al. (2025)	DS	–	–	–	–	–	Y	P	P	N	Y	N	N	N	N	2
[91]*	V. Spadari et al. (2025)	DS	–	–	–	–	–	Y	N	Y	N	P	P	N	N	P	2
[72]†*	K. Shaw et al. (2026)	DS	–	–	–	–	–	Y	N	N	N	P	P	N	N	P	1
[92]	P. Agarwal et al. (2026)	DS	–	–	–	–	–	P	P	P	N	P	N	N	P	P	0
Simulated quantum noise																	
[5]	F. Cirillo et al. (2026)	SN	–	–	–	–	–	N	P	P	N	Y	Y	N	N	P	2
[79]	A. Kuklinsky et al. (2024)	SN, HW	–	–	–	–	–	N	N	N	N	P	Y	Y	N	P	2
[71]†	A. Danasabe et al. (2026)	SN	–	–	–	–	–	N	P	P	N	Y	Y	N	N	N	2
[87]*	V. T. Venkateswarlu et al. (2026)	SN, FD	–	–	–	–	–	N	N	N	N	Y	Y	N	N	N	2

[75] †*	Y. Peng et al. (2024)	SN, HW	–	–	–	–	–	N	N	P	N	P	Y	Y	N	P	2
[10] *	A. Vyas et al. (2026)	SN, FD	–	–	–	–	–	P	P	P	N	P	Y	N	N	P	1
[94] *	I. Sabeshuly et al. (2025)	SN	–	–	–	–	–	N	N	N	N	P	Y	N	N	P	1
[82] *	W. Hammami et al. (2025)	SN	–	–	–	–	–	P	P	N	N	P	Y	N	N	P	1
[90] *	D. Abreu et al. (2025)	SN, HW (p)	–	–	–	–	–	P	N	N	N	P	Y	P	N	P	1
[93] *	N. Saxena et al. (2026)	SN	–	–	–	–	–	N	P	N	N	P	N	N	N	N	0
[74] †*	H. Suryotrisongko and Y. Musashi (2022)	SN	–	–	–	–	–	P	N	P	N	N	N	N	P	P	0
[88] *	D. Zhao et al. (2026)	SN, FD	–	–	–	–	–	N	N	N	N	N	P	N	N	P	0
Hardware execution only																	
[7] †	S. Badami (2026)	HW	–	–	–	–	–	N	P	P	P	Y	N	Y	Y	N	3
[6]	M. L. Cotrupi and B. R. Callahan (2025)	HW	–	–	–	–	–	N	N	P	N	Y	N	Y	N	P	2
[70] †*	A. Alomari and S. A. P. Kumar (2023)	HW	–	–	–	–	–	N	P	N	N	P	Y	Y	N	P	2
Yes (n = 33)			1	4	6	0	2	10	4	6	4	12	20	6	5	0	
Partial (n = 33)			6	0	2	1	0	9	12	14	4	18	4	2	5	23	
No (n = 33)			2	5	1	8	7	14	17	13	25	3	9	25	23	10	
N/A (n = 33)			24	24	24	24	24	0	0	0	0	0	0	0	0	0	

Yes / Partial / No without borderline (n = 26)	1/6/1	3/0/5	5/2/1	0/1/7	2/0/6	9/8/9	4/9/13	6/9/11	3/3/20	10/14/2	17/3/6	3/2/21	4/3/19	0/19/7	

Y = Yes, P = Partial, N = No, – = N/A (B1–B5 do not apply when no attack was run). Items as in Table 3: B1 threat model; B2 attack stronger than single-step FGSM; B3 budget reported and varied; B4 gradient-masking checks; B5 traffic constraints; B6 classical baseline under the same test; B7 baseline tuning or matched budget; B8 repeated runs with variance; B9 significance testing; B10 dataset issues; B11 metrics beyond accuracy under the test; B12 real quantum hardware; B13 code or data released; B14 conclusions within the evidence (codes adjudicated with the rule of Section III-C; the first appraisal gave Partial 14 and No 19, and 13/13 without borderline studies). Dim.: AE adversarial manipulation of inputs; PB poisoning or Byzantine; PX privacy or model extraction; SN simulated quantum noise; HW real hardware (HW(p) partial); DS distribution shift or unseen attacks; IN classical input noise or degradation; FD federated or distributed setting. Studies are grouped by the first dimension in the order AE/PX, DS, SN, HW. Yes = number of items coded Yes. † Borderline inclusion; * single appraisal of B1–B13 (B14 was appraised twice for every study).

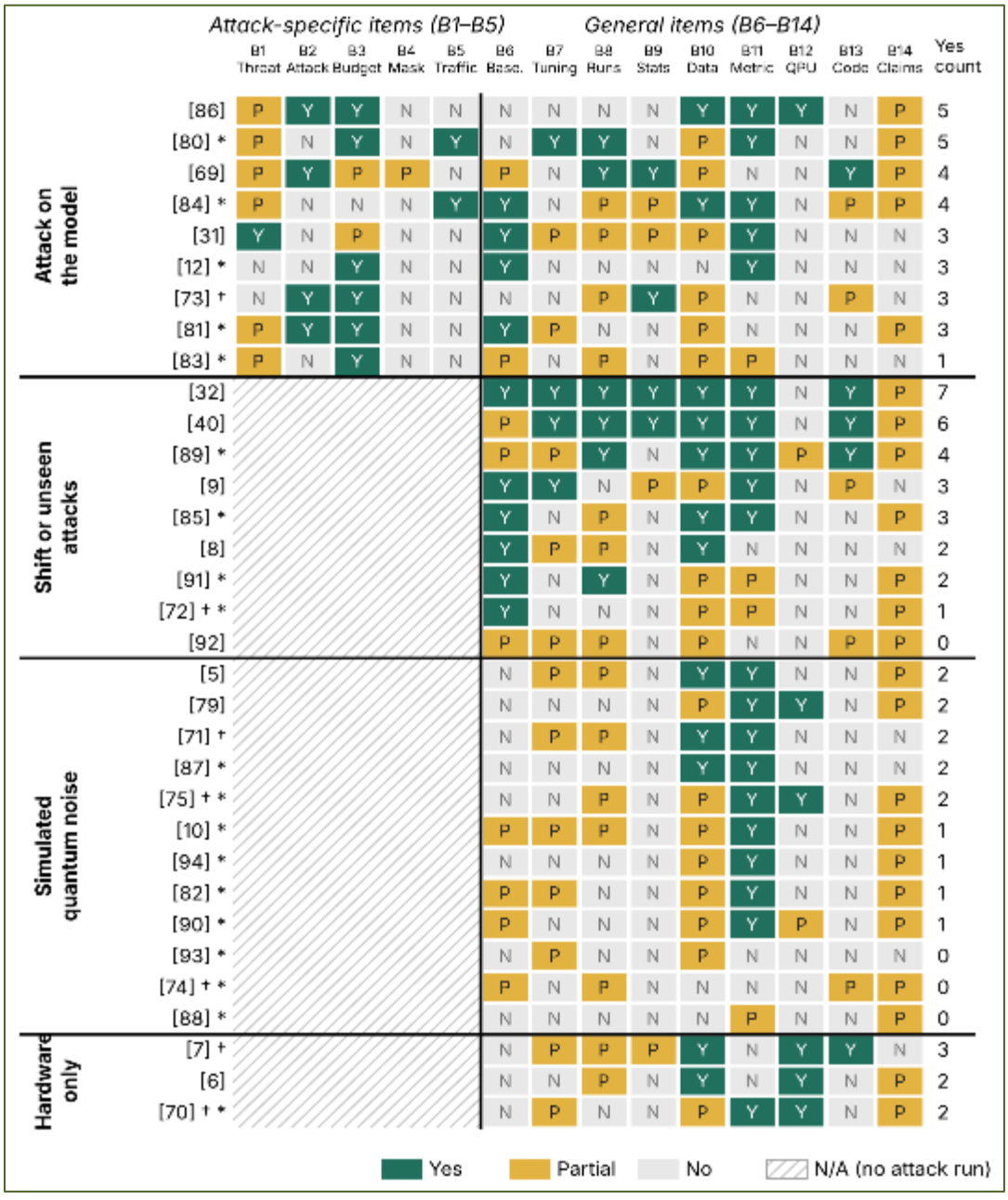


Figure 3: Heat map of the rubric codes (B1–B14) for the 33 included studies, grouped by the primary robustness dimension tested (attack on the model; distribution shift or unseen attacks; simulated quantum noise; hardware execution only) and sorted within each group by the number of items coded Yes (right). Hatched cells: N/A, no attack run. B14: adjudicated codes. † borderline inclusion; * single appraisal of B1–B13.

6.1 Attack Evaluation (B1–B5)

A threat model was stated in full in one attack study [31] and not at all in two [12], [73]. No study ran an attack ensemble or an attack adapted to its defence, and none checked gradient masking in full (B4: Partial 1, No 8); the partial case

treated PGD with exact density-matrix gradients as adaptive and left expectation over transformation (EOT) for shot-based deployment to future work [69]. This matters because defences evaluated with attempted adaptive attacks were still circumvented [52], an attack ensemble lowered the robust accuracy of all but one of more than 50 models [53], and two included defences relied on injected noise [69], [81], the stochastic kind whose robustness has often been traced to obfuscated gradients [51], [62]. Traffic constraints (B5) were respected only by the two fixed manipulations [80], [84], so accuracy under the optimised attacks bounds neither the vulnerability nor the robustness of a deployed detector [54], [64].

6.2 Is Quantum Noise a Defence?

Three attack studies tested noise as a defence, with conflicting results. For a QSVR detector, neither noise nor adversarial training improved robustness to PGD, and the attacks were crafted on the noiseless model [86]. For a stochastic QNN with depolarising noise, PGD $\ell\infty$ accuracy on NSL-KDD rose from 0.624 to 0.790 over seven seeds ($p = 0.044$), but the $\ell 2$ difference was not significant, and on UNSW-NB15 a classical MLP outperformed the noisy model even under white-box attack [69]. A third study credited "inherent quantum advantage" although its defence injected differential-privacy and depolarising noise and no gradient-masking check was reported [81]. The wider literature is equally divided, with noise reported as protective on hardware [95] but unreliable [15], or even helpful to attackers [17], [96]. The decisive test, an attack adapted to the noisy model with EOT [17], was run by no included study, so the claim that quantum noise protects network detectors is not supported.

6.3 Baselines and Attribution (B6, B7)

A classical baseline faced the same robustness test in 10 studies, seven of them shift studies; none of the 17 noise studies did so in full. Part of this gap is structural, but it left claims that hybrid designs mitigate noise [5] without any reference for how a classical pipeline would degrade. In one preprint, the headline drop of classical detectors from 95% to 47% came from a weaker transfer attack on a single seed [69]. Baseline tuning or a matched budget (B7) was documented in four studies, and in the best-controlled of them tuned classical models matched or exceeded the quantum models [40]. Encoding was not systematically examined as a robustness factor in the studies for which sufficient methodological details were available. However, the available evidence does not allow this conclusion to be extended to all studies, particularly those with reduced data extraction. Although attacks were applied to PCA inputs [69], [86], and the encoding norm is known to influence robustness [46], the contribution of encoding choices remains insufficiently investigated. Without such ablation studies, robustness differences cannot be confidently attributed to the quantum component of a model.

6.4 Statistics, Data and Reporting (B8–B13)

Variance over repeated runs was reported in full by six studies, and 25 studies did not test the differences they reported (B9). Even tested comparisons were fragile: one study printed $p = 1$ beside $t = -4.78$ [73], and in another only two advantages survived false-discovery-rate correction [40]. Several headline differences were smaller than run-to-run spread [8] or equalled one or two test records [87]. Twelve studies handled dataset issues in full (B10); one attack study still used KDD Cup 99 [12], and at least seven relied on CIC-IDS2017 or CSE-CIC-IDS2018, whose labelling errors destabilise published results [57], [58]. In one study every reported accuracy was below the 97.13% that a constant attack prediction would reach [87]. Twenty studies reported metrics beyond accuracy (B11), which exposed failures accuracy hid, such as false alarms rising from 7 to 21 of 43 benign records [87]; detection at a low false-positive rate was reported only in [40], and quantum-native robustness metrics were absent [66]. Hardware was used in 6 studies, never for an attack (B12), and code was released in full by 5 (B13).

6.5 Claims Against the Evidence (RQ3)

At abstract level, 26 of 45 assessed reports whose abstracts named a test (58%) were included, against 3 of 24 (12.5%) whose abstracts only claimed robustness; 18 of those 24 (75%) contained no robustness evaluation at all. A robustness claim that named no test was therefore a poor indicator that any test had been run.

At the full-text level, none of the 33 studies kept all conclusions fully within the scope of their empirical evidence (B14: Partial = 23, No = 10, Yes = 0). All ten studies coded as No had abstracts that explicitly mentioned a robustness test. Six recurrent types of evidence claim mismatch in the appraisal notes were reported. The main and most commonly encountered was making the claim a system is robust under a given parameter or situation that have not been tested in it (6 studies), such as claiming that the model is robust to noise even though no such noise level was used in testing [71], a kind of quantum-based robustness without a control (8), the fact that entanglement was credited as the cause for the defense while the attack was actually modifying the classical data before the quantum encoding [12]; mismatching like for like or selections dictated by unseen test data (11), e.g., the best out of seven backends was reported[90]; the missing or refuted statistical foundation (13); inconsistency of the text with the authors' own tables (8) [83], [91]; and the robustness concluded from output that responded minimally to the perturbation (3) [75], [87], [93].

6.6 Sensitivity Analyses

We left 26 studies after removing the seven borderline ones, and the pattern of results was still the same (Table 5, bottom row). Still, the number of hardware-studies was 3 out of 26 because three of the hardware studies were borderline. B14 still had no Yes-code, but 7 studies had "Partial" and 7"No". On applying a stricter version of first-appraisal codes, B14 changed into a situation of "Partial" 14 and No 19, that the number of studies coded as a "No is between 10 and 19" and the situation of no studies classified as "Yes" remains the same no matter which codes are used.

7 DISCUSSION

7.1 Principal Findings in Context

The central finding is not that QML network detectors are fragile, but that their robustness has rarely been tested in a way that could show it. Nine of the 33 studies attacked the detector, none checked for gradient masking in full, none combined an optimised attack with traffic constraints, and no study kept its conclusions fully within its evidence. Robustness in this literature was therefore mostly asserted or inferred from noise sweeps rather than demonstrated against an adversary.

These results agree with earlier syntheses and narrow them. The 25 of 33 studies (76%) without hardware execution fall within the simulator-only shares of 80.6% and 83.3% reported for hybrid intrusion detection [22] and threat intelligence [23]. The focus on input-level evasion measured by accuracy degradation reported for QML in general [24] appears here in a narrower form: only nine attack studies, mostly feature-space FGSM or PGD. Where noise was found to be an unreliable passive defence on image benchmarks [15], the three included studies that tested noise as a defence on network data did not settle the question either way, because none adapted the attack to the noise. The gap between claims and evidence parallels quantum learning for power systems, where no application study offered a matched-budget classical baseline [25]; in both domains the missing control is the same.

7.2 Interpreting the Pattern

Detection of a pattern of this kind in the field can be explained by features of three different types that may not have actually been tested, as here they are being presented as hypotheses only. The first is that many detectors were small simulated circuits on the use of a limited number of features, so that the easiest way of testing a detector was to perturbe its feature vector; however, the execution of an attack in the form of a traffic modification that uses the concept of space needs to have a feature-extraction pipeline that is the part of the model and not the quantum model. Second, since noise sweeps are relatively inexpensive to run in a simulated setting and generate a robustness curve without requiring the presence of an adversary, they might be the reason that simulated noise appeared as the most frequent dimension (17 studies vs single appearances of poisoning and privacy attacks). Third, the story of outperforming classical machine learning [36] gives room for good comparisons and the claims audit aligns with that: 18 of 24 assessed reports that claimed robustness without specifying which test they had used lacked a robustness evaluation.

The same failure modes, namely missing adaptive attacks [28], unrealistic threat models [3] and weak evaluation practice [4], were documented for classical learning-based security. QML detectors therefore inherit known problems rather than introduce new ones, with two additions specific to quantum models: sampled outputs that can obscure gradients, and encodings such as (2) that can discard information before classification.

7.3 Implications

For authors, the least costly corrections would change the reading of most included studies: perturbing classical baselines in the same way, repeating runs with statistical tests, and confining conclusions to what was tested (Section VIII). For reviewers and editors, a robustness claim for a quantum detector is a claim about a specific attack and a specific baseline and should be checked against both; Table 6 can serve as a review aid. For practitioners, the evidence does not yet justify choosing a QML detector over a tuned classical one for robustness reasons, since where a matched comparison was made, tuned classical models matched or exceeded the quantum models in aggregate detection [40]. Finally, the two appraisers agreed on B14 no better than chance before an explicit rule was applied ($\kappa = -0.06$), which suggests that judging whether conclusions exceed their evidence is hard to do reliably and should rest on decision rules fixed before coding.

8 A MINIMUM EVALUATION AND REPORTING CHECKLIST

Table 6 transforms each aspect of the rubrics into an explicit requirement for justifying the robustness of QML network detectors, plus she comes up with two new requirements that appraisal indicated were decisive: noise-aware attacks presented the noise as a defense, and disclosure of hardware execution completely transparently. The last column reports the number of included papers that already satisfied each criterion.

Table 6: Minimum Evaluation and Reporting Checklist for Robustness Claims About QML Network Detectors

No.	Requirement	Applies to	Derived from	Met in full (Table 5)	Grounding sources
Threat and attack (studies that attack the detector)					
C1	State a threat model: the attacker's goal, knowledge and capability, and the point of the pipeline where the attacker acts (traffic, features, encoding, circuit, training data or federation).	Attack	B1	1 of 9	[3], [28], [59]
C2	Run at least one iterative attack in addition to single-step FGSM and, for any defence, an attack adapted to that defence.	Attack	B2	4 of 9	[52], [53], [60], [61]
C3	Report the threat norm and budget, and sweep the budget so that robust performance is given as a function of it.	Attack	B3	6 of 9	[28]
C4	Report checks for gradient masking, for example that accuracy falls to chance at an unbounded budget and that transfer or black-box attacks are not stronger than white-box ones; for stochastic, noisy or shot-sampled outputs, attack with expectation over transformation (EOT).	Attack	B4	0 of 9	[17], [51], [53], [62]
C5	Make the perturbation realisable on the network, by applying it to traffic or by enforcing domain constraints on the features, and optimise it against the detector.	Attack	B5	2 of 9 (0 with B2 Yes)	[29], [54], [63]– [65]
Noise and hardware					
C6	Specify each noise channel, its strength and where it acts (training, inference or both), and sweep the strength. Where noise is presented as a defence, attack the noisy model with a noise-aware attack rather than transferring attacks crafted on the noiseless model.	Noise	Sections V-B, VI-B (no rubric item)	0 of 17 adapted the attack to the noise	[17]
C7	Report the device, qubit count, shots, number of samples and number of runs, and state which experiments ran on hardware; where training and deployment backends may differ, evaluate transfer across backends.	Hardware	B12; Section VI-D	6 of 33 used hardware; 0 attacked on it	[33], [67]
Comparison and attribution (all studies)					
C8	Subject the classical comparators to the same attack, shift or input-level perturbation as the quantum model.	All	B6	10 of 33 (0 of 17 noise studies)	[37], [39]
C9	Tune the classical baselines with matched inputs and budget, and ablate the encoding and classical preprocessing before attributing a robustness difference to the quantum part of the model.	All	B7; Section VI-C	4 of 33	[25], [37], [46]
Statistics, data and metrics (all studies)					
C10	Repeat runs over seeds (and repeated jobs on hardware), report the variance, and report the test-set size so that the resolution of each metric is known.	All	B8	6 of 33	—
C11	Test the differences that are claimed with tests suited to classifier comparison, corrected for multiple comparisons.	All	B9	4 of 33	[55]

C12	Avoid KDD Cup 99, address the documented issues of the benchmark used, and report class balance with the accuracy of a constant predictor; for claims about shift, test across datasets.	All	B10	12 of 33 (2 of 10 shift studies across datasets)	[56]–[58]
C13	Report metrics beyond accuracy under the test: per-class recall, false-positive rate on benign traffic, detection at a fixed low false-positive rate, and attack success rate for attacks.	All	B11	20 of 33	[24], [66]
Reporting and claims (all studies)					
C14	Release code, configurations, seeds, data splits and noise models.	All	B13	5 of 33	—
C15	Name the robustness test in the abstract, and restrict conclusions to the dimension, model, dataset and backend that were tested; do not attribute robustness to a quantum property without the controls of C8 and C9.	All	B14	0 of 33	[28], [54]

Met in full: number of included studies coded Yes on the source item in Table 5; attack items (C1–C5) are counted over the nine studies that attacked the detector. C6 has no rubric item; its count is from Section VI-B. Grounding sources are those of Table 3, with [17], [33], [46] and [67] added for the requirements on noise, hardware and encoding.

No included study satisfied all checklist requirements, and no attack-focused study met all five-core attack-evaluation criteria (C1–C5) simultaneously. Three requirements could have substantially improved the assessment of most studies with relatively limited additional effort: applying equivalent perturbations to classical comparators (C8), repeating experimental runs and statistically testing differences (C10 and C11), and restricting conclusions to the conditions and dimensions actually evaluated (C15). Although the attack-specific requirements involve greater experimental effort, their absence means that studies may primarily demonstrate the effectiveness of the attack rather than the robustness of the detection model [51]– [53].

For future appraisals, B14 and C15 should be assessed using explicit decision rules established before coding to improve consistency and inter-rater agreement. The checklist was developed by the authors and has not yet undergone pilot testing.

9 RESEARCH GAPS AND ROADMAP

Fig. 4 crosses the robustness dimensions with the rigour items that limited the conclusions. Several dimensions were absent: no study tested a backdoor, an attack on hardware, an attack adapted to a noisy model, an attack both optimised and realisable, or a certified guarantee, although each has a basis outside network security [16], [17], [26], [45], [47], [97], [98]. Poisoning and privacy attacks were evaluated once each, and poisoning was named in only four map abstracts, so their scarcity is not merely an artefact of access. The most studied dimensions lacked controls: no noise study perturbed a classical baseline in the same way, the hardware studies had no same-test baseline, variance over repeated runs or significance testing, and the federated studies met none of the four rigour items in Fig. 4 in full. No QML detector of attacks on QKD links evaluated its own robustness [99], [100], and none faced adversarial traffic generated by quantum models, although quantum attacks have already evaded classical detectors [18], [101].

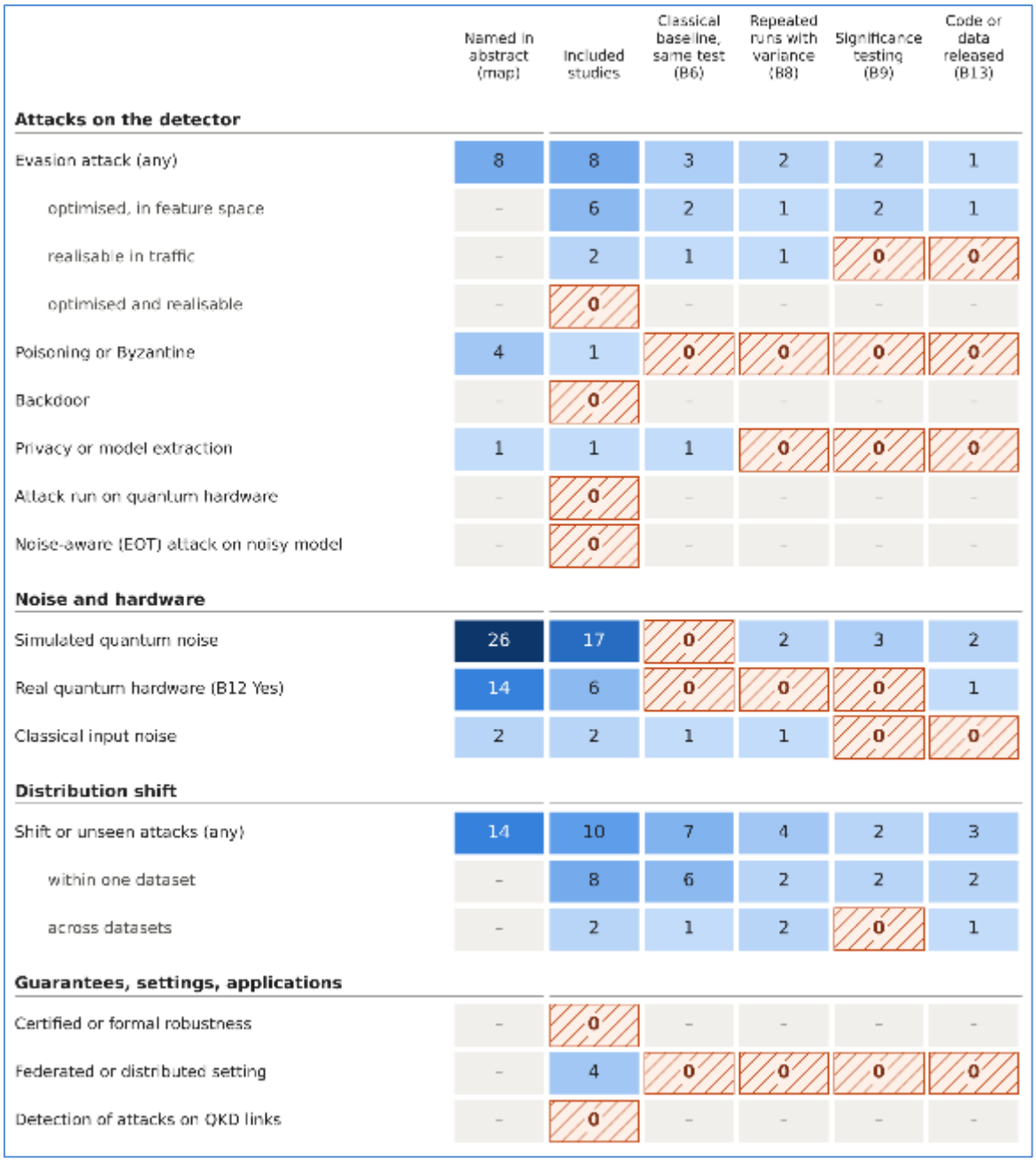


Figure 4: Gap map of robustness evaluation in QML network detectors. First column: number of the 76 map abstracts that named a test and named this type (types coded by R1; an abstract could name several). Second column: number of the 33 included studies that evaluated the dimension (Table5, Dim.; hardware is B12 Yes). Other columns: number of those studies coded Yes on rubric items B6, B8, B9 and B13. Hatched cells with 0: gap; grey cells with –: not coded at that level, or no study to count. Indented rows are subsets of the row above. QKD links: two QML detectors of attacks on QKD links were assessed at full text and excluded for lack of a robustness test [99], [100].

In the short term, published QML detectors should be re-evaluated under the checklist: adaptive attacks with gradient-masking checks, equally perturbed classical baselines, repeated runs with tests and false-positive rates on benign traffic. This needs no hardware and would provide the standardised protocols that earlier reviews requested [19], [24]. In the medium term, attacks should be optimised and realisable, using constraint-aware methods from classical NIDS research [29], [63], [64]; poisoning, backdoors and Byzantine clients need evaluation in federated settings [26], [97]; and shift should be tested across datasets and backends [67], with quantum-native metrics [66] and encoding ablations [46]. In the long term, certified bounds [16], [45], [47], [98] must be extended to constrained network features, and hardware validation [22], [24] must reach a scale that supports statistics.

10 THREATS TO VALIDITY

The evidence base is small and heterogeneous, so results were counted rather than pooled and a single study can shift a count. Many studies relied on dated or flawed benchmarks [56]–[58], incomplete reporting means that some No codes may reflect reporting rather than practice, nine studies were read as preprints or short versions, and favourable results may be over-represented [36].

The review process has its own limits. Only two databases were searched, for English-language records up to 29 September 2026. Screening relied on AI tools: 1,657 of 2,208 screened records rest on AI decisions alone, whereas rapid-review guidance asks for dual human screening of a share of abstracts and of all exclusions [102], and single screening misses a median of 5% of studies [103]. The robustness criterion was added after the records were known, its validation sample was stopped early, and the human re-check returned 104 of 243 exclusions to full text, showing that AI coding of this criterion was unreliable on its own. The two-tier design was a post hoc change, and most map fields were coded automatically. Extraction was AI-assisted and checked by humans [Tariq Adnan Fadil and Authors], the rubric is new, 15 studies carry a single appraisal, and the B14 rule was written after the disagreements had been seen. Finally, 128 of 204 reports could not be retrieved, so the second tier reflects access and may favour open-access work.

11 CONCLUSION

This review mapped 363 records on QML detectors for network security and appraised 33 studies. Robustness was evaluated mainly as simulated quantum noise and as distribution shift within single benchmarks, almost always on simulators; attacks on the detector appeared in nine studies, and backdoors and certified robustness in none. The evaluations fell short of published guidance on most rubric items, and 25 studies did not test the differences they reported. Robustness claims rarely held up: 18 of 24 assessed reports that claimed robustness without naming a test contained no robustness evaluation, and no included study kept its conclusions fully within its evidence. The checklist in Table 6 and the gap map in Fig. 4 set out the minimum evidence required. Until such evaluations are reported, the claim that QML makes network detectors more robust remains untested rather than refuted.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The authors thank Tariq Adnan Fadil, [Ministry of Higher Education and Scientific Research Iraq, Studies, Planning and Follow-up Directorate], for the human screening and stopping-rule checks, the independent second appraisal of the included studies and the blind second appraisal of item B14.

Use of AI tools. Claude Opus 5.5 (Anthropic) was used for the first-pass screening of all titles and abstracts (R1) and for blind AI re-screens of stratified samples; for retrieving open-access full texts and extracting study data, each value backed by verbatim quotes that were checked against the source text by script; for the first-pass coding of the appraisal rubric. ChatGPT (OpenAI; model version not recorded) screened a random sample of 441 records as the second screener (R2) on 20 October 2026. The screening checks, the stopping-rule checks and the second appraisal were carried out by a human reviewer without an AI tool, and the disagreements in the appraisal were adjudicated by the first author without an AI tool. The authors take full responsibility for the content of this article.

Disclosure of conflict of interest

The author declares that there are no conflicts of interest

Data, Code and Materials Availability

The review protocol and amendment log, the supplementary methods with the full search strings, Tables S1–S4, the screening workbook and the evaluation matrix are openly available on Zenodo (https://doi.org/10.5281/zenodo.24146315) under a CC BY 4.0 licence, together with the quote-verification script. Database abstracts and the full texts of the reviewed articles are not redistributed for copyright reasons; record identifiers, DOIs and all codes are provided.

REFERENCES

[1] D. Han et al., "Evaluating and Improving Adversarial Robustness of Machine Learning-Based Network Intrusion Detectors," IEEE Journal on Selected Areas in Communications, 2021. https://ieeexplore.ieee.org/document/9448103. (arXiv:2005.07519).

[2] A. Alotaibi et al., "Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense," Future Internet, vol. 15, no. 2, Art. no. 62, 2023. https://doi.org/10.3390/fi15020062.

[3] G. Apruzzese et al., "Modeling Realistic Adversarial Attacks against Network Intrusion Detection Systems," Digital Threats: Research and Practice, 2021. https://doi.org/10.1145/3469659. (arXiv:2106.09380)

[4] D. Arp et al., "Pitfalls in Machine Learning for Computer Security," Communications of the ACM, 2024. https://doi.org/10.1145/3643456.

[5] F. Cirillo, C. Esposito, and J. Taek Seo, "Benchmarking quantum machine learning methods for intrusion detection on noisy quantum computers," Quantum Machine Intelligence, vol. 8, no. 1, Art. no. 27, 2026. https://doi.org/10.1007/s42484-026-00379-4.

[6] M. L. Cotrupi and B. R. Callahan, "Towards Practical Quantum Kernels for Network Intrusion Detection," Proceedings of the AAAI Symposium Series, vol. 7, no. 1, pp. 339–342, 2025. https://ojs.aaai.org/index.php/AAAI-SS/article/view/36903.

[7] S. Badami, "Hardware-Agnostic Quantum Kernel Feature Mapping for Anomaly Detection in Critical Infrastructure: A Cross-Testbed Validation on NISQ Processors," IEEE Access, vol. 14, pp. 49642–49654, 2026. https://doi.org/10.1109/ACCESS.2026.3679234.

[8] Z. Wang, K. W. Fok, and V. L. L. Thing, "Network attack traffic detection with hybrid quantum-enhanced convolution neural network," Quantum Machine Intelligence, vol. 7, no. 1, Art. no. 50, 2025. https://doi.org/10.1007/s42484-025-00278-0.

[9] W. A. Malwi, F. Asiri, and M. S. Khan, "Quantum–Enhanced Intrusion Detection Using Quantum Circuit Born Machines for Zero-Day Attack Detection," Computers, Materials and Continua, vol. 88, no. 1, Art. no. 36, 2026. https://doi.org/10.32604/cmc.2026.075326.

[10] A. Vyas, R.-H. Hwang, P.-C. Lin, and M. Tripathi, "Communication-Efficient Quantum Federated Learning for Privacy-Preserving IIoT Network Intrusion Detection," in IEEE INFOCOM 2026 Workshops: Quantum Networked Applications and Protocols (QuNAP), 2026. https://doi.org/10.1109/INFOCOM59046.2026.11571554.

[11] M. T. West et al., "Towards quantum enhanced adversarial robustness in machine learning," Nature Machine Intelligence, vol. 5, no. 6, pp. 581–589, 2023. https://doi.org/10.1038/s42256-023-00661-1.

[12] H. Qi, J. Wang, and Y. Cui, "VQCNN: variational quantum convolutional neural networks based on quantum filters and fully connected layers," Computer Journal, vol. 67, no. 10, pp. 2970–2983, 2024. https://doi.org/10.1093/comjnl/bxae062.

[13] S. Lu, L.-M. Duan, and D.-L. Deng, "Quantum adversarial machine learning," Physical Review Research, vol. 2, no. 3, Art. no. 033212, 2020. https://doi.org/10.1103/PhysRevResearch.2.033212.

[14] F. Marchiori and M. Conti, "ATTAQ: Adversarial Robustness of Quantum Machine Learning," in 2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), pp. 200–207, 2025. https://doi.org/10.1109/DSN-W65791.2025.00062.

[15] S. R. Nowmi, J. Lopez, M. M. A. Imon, S. Pouryousef, and M. S. Rahman, "SoK: Critical Evaluation of Quantum Machine Learning for Adversarial Robustness," in Proc. IEEE Symposium on Security and Privacy, pp. 19–38, 2026. https://doi.org/10.1109/SP63933.2026.00137.

[16] Y. Du et al., "Quantum noise protects quantum classifiers against adversaries," Physical Review Research, vol. 3, no. 2, Art. no. 023153, 2021. https://doi.org/10.1103/PhysRevResearch.3.023153.

[17] H.-R. Wang et al., "Decoupling the Dual Impact of NISQ Noise on Quantum Adversarial Robustness," Entropy, vol. 28, no. 7, Art. no. 719, 2026. https://doi.org/10.3390/e28070719.

[18] I. P. Pathak, N. Yilmazer, and D. Pankajkumar Patel, "Quantum GAN-Based Adversarial Attack Generation for Network Intrusion Detection System," in 2025 ICCST - International Carnahan Conference on Security Technology, ICCST 2025, 2025. https://doi.org/10.1109/ICCST63435.2025.11295591.

[19] A. Kaissar, A. B. Nassif, and A. Bouridane, "Enhancing Network Intrusion Detection with Quantum Machine Learning: A Comprehensive Survey of Methods, Metrics, and Applications," Future Internet, vol. 18, no. 5, Art. no. 234, 2026. https://doi.org/10.3390/fi18050234.

[20] O. Faker and N. E. Cagiltay, "Quantum Machine Learning in Intrusion Detection Systems: A Systematic Mapping Study," in Lecture Notes in Networks and Systems, vol. 817, pp. 99–113, 2024. https://doi.org/10.1007/978-981-99-7886-1_9.

[21] O. K. Nicesio, A. G. Leal, and V. L. Gava, "Quantum Machine Learning for Network Intrusion Detection Systems, a Systematic Literature Review," in 2023 IEEE 2nd International Conference on AI in Cybersecurity, ICAIC 2023, 2023. https://doi.org/10.1109/ICAIC57335.2023.10044125.

[22] N. Louise and V. Kavitha, "Hybrid Quantum-Classical Machine Learning for Cybersecurity Threat Detection and Intrusion Prediction: A Systematic Review," in 2026 4th International Conference on Artificial Intelligence and Machine Learning Applications: Healthcare and Internet of Things, AIMLA 2026, 2026. https://doi.org/10.1109/AIMLA67915.2026.11522502.

[23] P. Rezaei, B. Rahgozar, M. S. Khan, M. Yadollahi, and M. Saberikamarposhti, "Quantum Machine Learning for Cyber Threat Intelligence: A Scoping Review of Current Capabilities and Future Directions," in Proc. International Colloquium on Signal Processing and Its Applications, CSPA, pp. 260–265, 2026. https://doi.org/10.1109/CSPA68262.2026.11517902.

[24] Y. A. Kustiawan and K. I. Ghauth, "Adversarial Robustness in Quantum Machine Learning: A Scoping Review," Computers, vol. 15, no. 4, Art. no. 233, 2026. https://doi.org/10.3390/computers15040233.

[25] L. Tightiz and E. Oh, "Quantum Learning in Modern Power Systems: A Critical Appraisal of Current Evidence and Deployment Barriers," Journal of Modern Power Systems and Clean Energy, early access, 2026. https://doi.org/10.35833/MPCE.2026.000496.

[26] V.-L. Nguyen, L.-H. Nguyen, R.-H. Hwang, B. Canberk, and T. Q. Duong, "Quantum Machine Learning for 6G Network Intelligence and Adversarial Threats," IEEE Communications Standards Magazine, vol. 9, no. 3, pp. 40–48, 2025. https://doi.org/10.1109/MCOMSTD.2025.3575261.

[27] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, Art. no. n71, 2021. https://doi.org/10.1136/bmj.n71.

[28] N. Carlini et al., "On Evaluating Adversarial Robustness," arXiv:1902.06705, 2019. https://arxiv.org/abs/1902.06705.

[29] J. Vitorino et al., "SoK: Realistic adversarial attacks and defenses for intelligent network intrusion detection," Computers & Security, 2023. https://www.sciencedirect.com/science/article/pii/S0167404823003437. (arXiv:2308.06819)

[30] V. Havlíček et al., "Supervised learning with quantum-enhanced feature spaces," Nature, vol. 567, p. 209, 2019. https://doi.org/10.1038/s41586-019-0980-2.

[31] S. Rawas, M. Al Saleh, A. D. Samala, and A. Fortuna, "Beyond encryption: quantum-enhanced behavioral security for the post-quantum era," Applied Computing and Informatics, pp. 1–15, 2026. https://doi.org/10.1108/ACI-03-2026-0142.

[32] A. Thomos and T. Andronikos, "Evaluation of Variational Quantum Classifiers (VQC) for Cyberattack Detection in the NISQ Era," Journal of Sensor and Actuator Networks, vol. 15, no. 4, Art. no. 67, 2026. https://doi.org/10.3390/jsan15040067.

[33] J. Preskill, "Quantum Computing in the NISQ era and beyond," Quantum, vol. 2, p. 79, 2018. https://doi.org/10.22331/q-2018-08-06-79.

[34] J. McClean et al., "Barren plateaus in quantum neural network training landscapes," Nature Communications, 2018. https://doi.org/10.1038/s41467-018-07090-4.

[35] S. Wang et al., "Noise-induced barren plateaus in variational quantum algorithms," Nature Communications, 2021. https://doi.org/10.1038/s41467-021-27045-6.

[36] M. Schuld and N. Killoran, "Is Quantum Advantage the Right Goal for Quantum Machine Learning?" PRX Quantum, vol. 3, Art. no. 030101, 2022. https://doi.org/10.1103/PRXQuantum.3.030101.

[37] J. Bowles et al., "Better than classical? The subtle art of benchmarking quantum machine learning models," arXiv:2403.07059, 2024. https://arxiv.org/abs/2403.07059.

[38] H.-Y. Huang et al., "Power of data in quantum machine learning," Nature Communications, 2021. https://doi.org/10.1038/s41467-021-22539-9.

[39] F. J. Schreiber et al., "Classical surrogates for quantum learning models," Physical Review Letters, vol. 131, no. 10, Art. no. 100803, 2023. https://doi.org/10.1103/PhysRevLett.131.100803.

[40] S. A. Gillani, M. S. A. Baig, S. M. Shah, A. Ali, and H. Siddiqui, "How Quantum Is the Advantage? A Fair, Calibration- and Noise-Aware Benchmark and Attribution Audit of Quantum Machine Learning for Network Intrusion Detection," arXiv:2608.18155, 2026. https://arxiv.org/abs/2608.18155.

[41] N. Dowling et al., "Adversarial robustness guarantees for quantum classifiers," npj Quantum Information, 2025. https://doi.org/10.1038/s41534-025-01129-3. (arXiv:2405.10360)

[42] N. Liu and P. Wittek, "Vulnerability of quantum classification to adversarial perturbations," Physical Review A, vol. 101, no. 6, Art. no. 062331, 2020. https://doi.org/10.1103/PhysRevA.101.062331.

[43] W. Ren et al., "Experimental quantum adversarial learning with programmable superconducting qubits," Nature Computational Science, 2022. https://doi.org/10.1038/s43588-022-00351-9.

[44] G. Montalbano and L. Banchi, "Quantum adversarial learning for kernel methods," Quantum Machine Intelligence, vol. 7, no. 1, Art. no. 15, 2025. https://doi.org/10.1007/s42484-025-00238-8.

[45] M. Weber et al., "Optimal provable robustness of quantum classification via quantum hypothesis testing," npj Quantum Information, 2021. https://doi.org/10.1038/s41534-021-00410-5.

[46] J. Berberich, D. Fink, D. Pranjić, C. Tutschku, and C. Holm, "Training robust and generalizable quantum models," Physical Review Research, vol. 6, no. 4, Art. no. 043326, 2024. https://doi.org/10.1103/PhysRevResearch.6.043326.

[47] A. Saxena, T. Wollschlager, N. Franco, J. M. Lorenz, and S. Gunnemann, "Certifiably Robust Encoding Schemes," in Proc. IEEE Quantum Week 2024, QCE 2024, pp. 1571–1582, 2024. https://doi.org/10.1109/QCE60285.2024.00184.

[48] M. L. Rethlefsen et al., "PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews," Systematic Reviews, vol. 10, Art. no. 39, 2021. https://doi.org/10.1186/s13643-020-01542-z.

[49] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, pp. 37–46, 1960. https://doi.org/10.1177/001316446002000104.

[50] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159–174, 1977. https://pubmed.ncbi.nlm.nih.gov/843571/.

[51] A. Athalye et al., "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples," in Proc. Int. Conf. Machine Learning (ICML), PMLR vol. 80, 2018. https://proceedings.mlr.press/v80/athalye18a.html.

[52] F. Tramèr et al., "On Adaptive Attacks to Adversarial Example Defenses," in Advances in Neural Information Processing Systems (NeurIPS), 2020. https://proceedings.neurips.cc/paper/2020/hash/11f38f8ecd71867b42433548d1078e38-Abstract.html.

[53] F. Croce et al., "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in Proc. 37th Int. Conf. Machine Learning (ICML), PMLR vol. 119, 2020. https://proceedings.mlr.press/v119/croce20b.html.

[54] M. M. Hasan et al., "Adversarial attacks against network intrusion detection systems: Bridging the gap between theoretical vulnerabilities and practical constraints," Internet of Things, 2026. https://www.sciencedirect.com/science/article/pii/S2542660526001277.

[55] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006. https://jmlr.org/papers/v7/demsar06a.html.

[56] M. Tavallaee et al., "A detailed analysis of the KDD CUP 99 data set," in Proc. IEEE Symp. Computational Intelligence for Security and Defense Applications (CISDA), 2009. https://ieeexplore.ieee.org/document/5356528.

[57] G. Engelen et al., "Troubleshooting an Intrusion Detection Dataset: the CICIDS2017 Case Study," in Proc. IEEE Security and Privacy Workshops (SPW), 2021. https://ieeexplore.ieee.org/document/9474286.

[58] L. Liu et al., "Error Prevalence in NIDS datasets: A Case Study on CIC-IDS-2017 and CSE-CIC-IDS-2018," in Proc. IEEE Conf. Communications and Network Security (CNS), 2022. https://ieeexplore.ieee.org/document/9947235.

[59] P. Debus et al., "Entangled Threats: A Unified Kill Chain Model for Quantum Machine Learning Security," in Proc. IEEE Quantum Week 2025, QCE 2025, pp. 1653–1664, 2025. https://doi.org/10.1109/QCE65121.2025.00183.

[60] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learning Representations (ICLR), 2015. https://arxiv.org/abs/1412.6572.

[61] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. Int. Conf. Learning Representations (ICLR), 2018. https://arxiv.org/abs/1706.06083.

[62] N. Popovic et al., "Gradient Obfuscation Checklist Test Gives a False Sense of Security," arXiv:2206.01705, 2022. https://arxiv.org/abs/2206.01705.

[63] N. Alhussien et al., "Constraining Adversarial Attacks on Network Intrusion Detection Systems: Transferability and Defense Analysis," IEEE Transactions on Network and Service Management, 2024. https://doi.org/10.1109/TNSM.2024.3357316.

[64] M. Catillo et al., "Towards realistic problem-space adversarial attacks against machine learning in network intrusion detection," in Proc. 19th Int. Conf. Availability, Reliability and Security (ARES), 2024. https://doi.org/10.1145/3664476.3669974.

[65] F. Pierazzi, F. Pendlebury, J. Cortellazzi, and L. Cavallaro, "Intriguing properties of adversarial ML attacks in the problem space," in Proc. IEEE Symp. Security and Privacy (SP), 2020, pp. 1332–1349. https://doi.org/10.1109/SP40000.2020.00073.

[66] N. Innan, W. E. Maouaki, A. Marchisio, and M. Shafique, "Quantum Neural Networks Under Threat: Modeling Security Risks and Attack Vectors," in Communications in Computer and Information Science, vol. 2744, pp. 129–139, 2026. https://doi.org/10.1007/978-3-032-13855-2_12.

[67] V. Reers and M. Mausner, "A Threat-Model-Driven Robustness Benchmark for Quantum Machine Learning under Device Noise and Deployment Shift," in 2026 International Conference on Quantum Control, Computing and Learning, qCCL 2026, 2026. https://doi.org/10.1109/qCCL70331.2026.11668558.

[68] V. Sammartino, "Hybrid quantum Graph Neural Networks for robust botnet detection in modern IoT ecosystems," Future Generation Computer Systems, vol. 185, Art. no. 108650, 2026. https://doi.org/10.1016/j.future.2026.108650.

[69] G.-E. Filardo, "Decoherence as Defence and the Magnitude of Noise Regularisation: A Rigorous N-Qubit Theory of Stochastic Quantum Neural Networks for Adversarially Robust Network Intrusion Detection," arXiv:2606.24219, 2026. https://arxiv.org/abs/2606.24219.

[70] A. Alomari and S. A. P. Kumar, "DEQSVC: Dimensionality Reduction and Encoding Technique for Quantum Support Vector Classifier Approach to Detect DDoS Attacks," IEEE Access, vol. 11, pp. 110570–110581, 2023. https://doi.org/10.1109/ACCESS.2023.3322723.

[71] A. Danasabe, M. Afaq, A. H. El-Maleh, A. Khan, S. El-Ferik, and Z. Kaleem, "QATNet: A Lightweight Quantum-Classical Tabular Network for Low-Latency Intrusion Detection," IEEE Transactions on Quantum Engineering, vol. 7, Art. no. 3102015, 2026. https://doi.org/10.1109/TQE.2026.3674396.

[72] K. Shaw, S. Mishra, and I. Zelinka, "Agentic AI-enhanced quantum computing for cybersecurity: a new horizon in internet defense," Quantum Information Processing, vol. 25, no. 6, Art. no. 182, 2026. https://doi.org/10.1007/s11128-026-05161-w.

[73] H. Suryotrisongko, Y. Musashi, A. Tsuneda, and K. Sugitani, "Adversarial Robustness in Hybrid Quantum-Classical Deep Learning for Botnet DGA Detection," Journal of Information Processing, vol. 30, pp. 636–644, 2022. https://doi.org/10.2197/IPSJJIP.30.636.

[74] H. Suryotrisongko and Y. Musashi, "Evaluating hybrid quantum-classical deep learning for cybersecurity botnet DGA detection," Procedia Computer Science, vol. 197, pp. 223–229, 2022. https://doi.org/10.1016/j.procs.2021.12.135.

[75] Y. Peng, X. Li, Z. Liang, and Y. Wang, "HyQ2: A Hybrid Quantum Neural Network for NextG Vulnerability Detection," IEEE Transactions on Quantum Engineering, vol. 5, Art. no. 3103819, 2024. https://doi.org/10.1109/TQE.2024.3481280.

[76] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in Proc. Military Communications and Information Systems Conf. (MilCIS), 2015, pp. 1–6. https://doi.org/10.1109/MilCIS.2015.7348942.

[77] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. 4th Int. Conf. Information Systems Security and Privacy (ICISSP), 2018, pp. 108–116. https://doi.org/10.5220/0006639801080116.

[78] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: a new generation dataset of IoT and IIoT for data-driven intrusion detection systems," IEEE Access, vol. 8, pp. 165130–165150, 2020. https://doi.org/10.1109/ACCESS.2020.3022862.

[79] A. Kukliansky, M. Orescanin, C. Bollmann, and T. Huffmire, "Network Anomaly Detection Using Quantum Neural Networks on Noisy Quantum Computers," IEEE Transactions on Quantum Engineering, vol. 5, pp. 1–11, 2024. https://doi.org/10.1109/TQE.2024.3359574.

[80] T. Hammadia, A. M. Saber, M. Kassouf, D. Kundur, and M. Khalaf, "Optimal Quantum Variational Circuits for Detecting False Data Injection Attacks on Transmission Substations," IEEE Transactions on Industry Applications, early access, 2026. https://doi.org/10.1109/TIA.2026.3677243.

[81] H. M. Ngo, T. R. Jeter, J. T. Seo, and M. T. Thai, "QUPID: A Partitioned Quantum Neural Network for Anomaly Detection in Smart Grid," IEEE Transactions on Smart Grid, vol. 17, no. 6, pp. 6046–6056, 2026 (read as arXiv:2601.11500). https://doi.org/10.1109/TSG.2026.3705645.

[82] W. Hammami, S. Cherkaoui, and S. Wang, "Enhancing Network Anomaly Detection with Quantum GANs and Successive Data Injection for Multivariate Time Series," in Proc. 21st Int. Wireless Communications and Mobile Computing Conf. (IWCMC), pp. 1667–1672, 2025 (read as arXiv:2505.11631). https://doi.org/10.1109/IWCMC65282.2025.11059562.

[83] K. Addo, M. Kabeya, and E. E. Ojo, "Federated Quantum Machine Learning for Distributed Cybersecurity in Multi-Agent Energy Systems," Energies, vol. 18, no. 20, Art. no. 5418, 2025. https://doi.org/10.3390/en18205418.

[84] T. Ngobeni and B. Kabaso, "Hybrid Quantum–Classical Anomaly Detection for 5G Roaming Signalling with QKD/QSDC Support," Future Internet, vol. 18, no. 8, Art. no. 396, 2026. https://doi.org/10.3390/fi18080396.

[85] C. Rosa-Remedios, P. Caballero-Gil, and J. Molina-Gil, "Hybrid Quantum-Kernel and Quantum-Inspired Machine Learning for TDoS Early Warning in Critical Infrastructure," Computers, Materials and Continua, vol. 89, no. 1, Art. no. 11, 2026. https://doi.org/10.32604/cmc.2026.084352.

[86] K. Tscharke, M. Wendlinger, S. Issel, and P. Debus, "Anomaly Detection with Quantum SVR in the NISQ Era: Limits of Robustness to Noise and Adversarial Attacks," in International Conference on Agents and Artificial Intelligence, pp. 1114–1125, 2026. https://doi.org/10.5220/0014218600004052.

[87] V. T. Venkateswarlu et al., "A lightweight hybrid quantum classical federated learning model for anomaly detection in edge IoT networks," Journal of Theoretical and Applied Information Technology, vol. 104, no. 16, pp. 109–121, 2026. https://doi.org/10.5281/zenodo.23140824.

[88] D. Zhao, J. Meng, and X. Shang, "ϵ-DQML: A Framework for Distributed Quantum Machine Learning over Noises and Link Failures," in Proc. 2026 ACM SIGCOMM Workshop on Quantum Networks and Distributed Quantum Computing (QuNet), pp. 7–12, 2026. https://doi.org/10.1145/3833409.3833439.

[89] R. Verma and R. Bhatt, "QSAFE-Net: A Quantum Self-attention Flow Encoder Network for Cross-dataset Big Data Intrusion Detection in Cloud Environments," International Journal of Intelligent Engineering and Systems, vol. 19, no. 7, pp. 843–855, 2026. https://doi.org/10.22266/ijies2026.0731.48.

[90] D. Abreu, D. Moura, C. Esteve Rothenberg, and A. Abelém, "QuantumNetSec: Quantum Machine Learning for Network Security," International Journal of Network Management, vol. 35, no. 4, Art. no. e70018, 2025. https://doi.org/10.1002/nem.70018.

[91] V. Spadari, I. Guarino, D. Ciuonzo, and A. Pescapè, "Towards Network Intrusion Detection via Quantum Machine Learning: A Reality Check," in IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2025, 2025. https://doi.org/10.1109/INFOCOMWKSHPS65812.2025.11152772.

[92] P. Agarwal, R. K. Srivastava, D. K. Nishad, M. A. Mossie, and S. Gupta, "A hybrid SHA-256 and variational quantum classifier framework for simulation-based intrusion detection in SDN-IoT networks," Discover Internet of Things, vol. 6, no. 1, Art. no. 126, 2026. https://doi.org/10.1007/s43926-026-00474-9.

[93] N. Saxena, A. Nigam, B. K. Behera, and F. J. Seo, "Investigation of quantum machine learning algorithms in noisy environment for cybersecurity applications," AIP Advances, vol. 16, Art. no. 035226, 2026. https://doi.org/10.1063/5.0281142.

[94] I. Sabeshuly, A. Akzhalova, and S. B. Yahia, "Quantum-enhanced blockchain security: integrating quantum computing with network attack detection," Herald of the Kazakh-British Technical University, vol. 22, no. 3, pp. 123–133, 2025. https://doi.org/10.55452/1998-6688-2025-22-3-123-133.

[95] H.-F. Zhang et al., "Experimental robustness benchmarking of quantum neural networks on a superconducting quantum processor," Science China: Physics, Mechanics and Astronomy, vol. 69, no. 6, Art. no. 260315, 2026. https://doi.org/10.1007/s11433-025-2943-6.

[96] M. Pan, W. Liu, Y. Xu, S. Sun, and F. Yang, "Momentum-based quantum adversarial attack algorithm," Quantum Machine Intelligence, vol. 8, no. 1, Art. no. 62, 2026. https://doi.org/10.1007/s42484-026-00409-1.

[97] C. Chu, F. Chen, P. Richerme, and L. Jiang, "QDoor: Exploiting Approximate Synthesis for Backdoor Attacks in Quantum Neural Networks," in Proc. 2023 IEEE International Conference on Quantum Computing and Engineering, QCE 2023, pp. 1098–1106, 2023. https://doi.org/10.1109/QCE57702.2023.00124.

[98] J. Guan, W. Fang, and M. Ying, "Robustness verification of quantum classifiers," in Proc. 33rd Int. Conf. Computer Aided Verification (CAV), Lecture Notes in Computer Science, vol. 12759, 2021, pp. 151–174. https://doi.org/10.1007/978-3-030-81685-8_7.

[99] X. He, J. Xiao, and X. Lyu, "A Hybrid Quantum–Classical Neural Network Framework for the Detection of Quantum Hacking Attacks in CVQKD," Applied Sciences, vol. 16, no. 6, Art. no. 2793, 2026. https://doi.org/10.3390/app16062793.

[100] A. Al-Kuwari, N. Mohamed, S. Al-Kuwari, A. Farouk, and B. K. Behera, "Resisting Quantum Key Distribution Attacks Using Quantum Machine Learning," IET Quantum Communication, vol. 7, no. 1, Art. no. e70028, 2026. https://doi.org/10.1049/qtc2.70028.

[101] G. Sipani, A. J. Jose, S. Ravi, Y. Cao, and L. Watkins, "Quantum Computing: An Emerging or Immediate Threat to Classical AI-Based Security Systems?," in 2026 IEEE 16th Annual Computing and Communication Workshop and Conference, CCWC 2026, pp. 526–532, 2026. https://doi.org/10.1109/CCWC67433.2026.11393885.

[102] C. Garritty et al., "Cochrane Rapid Reviews. Interim guidance from the Cochrane Rapid Reviews Methods Group," Cochrane Rapid Reviews Methods Group, Mar. 2020. https://methods.cochrane.org/sites/methods.cochrane.org.rapidreviews/files/uploads/cochrane_rr_-_guidance-23mar2020-final.pdf.

[103] S. Waffenschmidt, M. Knelangen, W. Sieben, S. Bühn, and D. Pieper, "Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review," BMC Medical Research Methodology, vol. 19, Art. no. 132, 2019, doi: 10.1186/s12874-019-0782-0. https://bmcmedresmethodol.biomedcentral.com/track/pdf/10.1186/s12874-019-0782-0.