

Authenticity, integrity and provenance assurance across the cyber stack: A systematic review of detection, watermarking, authentication and cryptographic mechanisms

Kolawole Olamide [Michael](#)^{a,*}, Lebo Norabel [Ikaka](#)^a, Nathaniel Etim [Effiom](#)^a, Murphy Reje-Okim [Mbeh](#)^a, Faith Bassey-Ita [Bassey](#)^a, Bitrus Levi [Garki](#)^a, Oyui Daniel [Lawrence](#)^a, Samuel Okon [Essang](#)^b

^aDepartment of Mathematics and Computer Science, Arthur Jarvis University, Akpabuyo, Cross River State, Nigeria

^bMathematics Programme, Department of Physical Sciences, Landmark University, Omu-Aran, Kwara State, Nigeria

ARTICLE INFO

Article history:

Received: 31 July 2026

Received in revised form: 14 September 2026

Accepted: 25 September 2026

Available online: 19 October 2026

Keywords: Systematic review, PRISMA 2020, Integrity assurance, Provenance verification, Repackaging detection

DOI: [10.61298/rans.2026.4.2.416](https://doi.org/10.61298/rans.2026.4.2.416)

ABSTRACT

Assurance of authenticity, integrity and provenance recurs across the computing stack, but relevant research communities often publish in isolation. This protocol-driven systematic review appraises a bounded contemporary evidence base across five assurance layers—application, media, network, protocol and implementation—to identify cross-layer regularities and evidential asymmetries. Following PRISMA 2020, we enumerated peer-reviewed research published in Cyber Security and Applications, Volume 4 (2026), as available on 15 September 2026. Nine records were identified; four were excluded at title/abstract screening under the pre-specified eligibility criteria, and five satisfied all inclusion criteria at full-text assessment. Statistical heterogeneity precluded meta-analysis, so a narrative synthesis was undertaken. The included studies span application artefacts, media objects, physical-network devices, cryptographic protocols and micro-architectural implementations. Four use machine learning and one supplies security proofs with formal verification. Reported outcomes include watermark accuracy of 99.91%, spoofing-detection accuracy of 94.18%, total signcryption cost of 887 ms for ten receivers with linearly growing transmission cost, and recovery of a masked AES key within approximately 60 attack traces. Methodological quality ranged from 81.3% to 93.8%. Across layers, assurance mechanisms converge on an invariance principle: they seek representations stable under benign transformation but sensitive to adversarial modification. The evidence also shows an asymmetry between empirical and provable assurance and uneven benchmark maturity across layers. The review was not prospectively registered; the protocol is reproduced in Section 2 and Appendices A–C.

© 2026 The Author(s). Production and Hosting by FLAYOO Publishing House LTD on Behalf of the Nigerian Society of Physical Sciences (NSPS). Peer review under the responsibility of NSPS. This is an open access article under the terms of the [Creative Commons Attribution 4.0 International license](#). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

1. INTRODUCTION

1.1. RATIONALE

A single question recurs, in different vocabularies, at every level of a modern computing system: is this artefact what it claims to be, and has it been altered since it left the hands of the party that vouched for it? A user installing a mobile application asks it of the application package. A broadcaster releasing synthetic

*Corresponding Author Tel. No.: +234-806-111-7099.

e-mail: olamidekom@gmail.com (Kolawole Olamide Michael)

speech asks it of the audio object. A utility receiving a consumption report asks it of the meter that transmitted it. A recipient of a broadcast ciphertext asks it of the sender. An evaluator certifying a smart card asks it of the silicon implementation itself. In each case the underlying concern is identical – authenticity, integrity and provenance – even though the objects, the threat surfaces and the mathematical apparatus differ enormously.

The research communities that address this concern have nevertheless developed along largely separate trajectories. Software-artefact analysis has produced a substantial literature on clone and repackaging detection; multimedia forensics has produced an equally substantial literature on watermarking and content authentication; wireless-network security has produced physical-layer authentication and intrusion detection; theoretical cryptography has produced signcryption and its many refinements; and implementation security has produced side-channel analysis and the countermeasures it motivates. These literatures cite one another rarely. The consequence is that structural regularities which are visible only when the layers are viewed together – for instance, the recurring trade-off between the abstraction level of a feature representation and its resilience to adversarial transformation – go largely unremarked.

A further consequence is methodological. Each community has evolved its own conventions for what constitutes adequate evidence. In provable cryptography, a security claim is expected to be discharged by a reduction to a hard problem under a clearly stated adversary model. In learning-based detection, the corresponding expectation is a benchmark comparison against re-trained baselines. Neither convention is wrong, but the divergence means that claims of comparable strength across the layers are not, in fact, supported by evidence of comparable strength. Making that asymmetry explicit is a precondition for reasoning coherently about the security of systems that span several layers at once – which is to say, about almost all deployed systems.

This review responds to that situation. It takes a bounded, contemporary and deliberately heterogeneous body of primary research and subjects it to a single, uniform appraisal and synthesis procedure, so that the mechanisms, the evidence and the limitations at each layer can be compared on common terms. The five included studies constituting the bounded evidence base are cited explicitly in the synthesis [1–5].

1.2. DESCRIPTION OF THE PHENOMENON OF INTEREST

For the purposes of this review, an assurance mechanism is defined as any technique whose purpose is to establish, verify or contest a claim about the authenticity, integrity or provenance of a digital artefact, message, device or computation. The definition is deliberately framed to admit three functionally distinct classes of contribution.

- Detective mechanisms observe an artefact after the fact and decide whether it has been tampered with or illegitimately derived from another artefact. Repackaging detection and intrusion detection are of this class.
- Constructive mechanisms embed verifiable structure into an artefact in advance, so that later verification becomes possible. Watermarking and signcryption are of this class.

- Evaluative mechanisms deliberately attack an artefact in order to quantify the assurance it actually provides, as distinct from the assurance it is claimed to provide. Side-channel analysis is of this class.

Admitting all three classes is not a loosening of scope but a requirement of it. An assurance ecosystem in which only defensive contributions were surveyed would systematically overstate the security of the whole, because the evaluative literature is precisely the literature that establishes the upper bound on what the defensive literature can deliver. The inclusion of an offensive study alongside four defensive ones is therefore a deliberate design decision, and the synthesis in Section 4 treats the relationship between the two as a substantive finding rather than an inconvenience.

1.3. REVIEW QUESTIONS AND OBJECTIVES

The review was designed to answer five questions, each of which maps onto a distinct extraction domain and a distinct results subsection:

1. RQ1 – What adversary models and threat capabilities are assumed by contemporary assurance mechanisms, and how completely are those models specified?
2. RQ2 – What technical mechanisms and feature representations are deployed, and at what level of abstraction do they operate?
3. RQ3 – What datasets, benchmarks and evaluation metrics are used, and to what extent are these standardised within and across layers?
4. RQ4 – What levels of performance, cost and overhead are actually demonstrated, and are the comparisons on which those claims rest methodologically sound?
5. RQ5 – What limitations do the primary authors concede, and what unresolved problems does the aggregate evidence identify?

The corresponding objectives were: to identify and screen the evidence base against pre-specified criteria; to extract structured data on a uniform form; to appraise methodological quality on a uniform rubric; to synthesise the evidence thematically across the five questions; and to derive from the identified gaps a prioritised research agenda.

1.4. CONTRIBUTION AND STRUCTURE OF THIS REVIEW

Four contributions are claimed. First, the review supplies what is, to the author's knowledge, the first uniform appraisal of assurance mechanisms spanning the application, media, network, protocol and implementation layers within a single protocol-driven synthesis. Second, it introduces an eight-domain appraisal rubric that is applicable to both empirical and provable-security contributions, and reports its application transparently at the item level. Third, it makes explicit a cross-layer invariance principle that organises the technical findings of all five included studies. Fourth, it derives a research agenda that is traceable, item by item, to specific gaps in the appraised evidence rather than to general exhortation.

Table 1. Eligibility criteria specified a priori (adapted PICOS framework).

Element	Inclusion criteria	Exclusion criteria
Population (protected asset)	Any digital artefact, message, device, computation or communication channel for which an authenticity, integrity or provenance claim can be made – including software packages, media objects, networked embedded devices, transmitted ciphertexts and cryptographic implementations.	Studies concerned exclusively with availability (e.g. denial-of-service mitigation) or exclusively with confidentiality where no authenticity or integrity claim is at issue.
Intervention (mechanism)	A detective, constructive or evaluative assurance mechanism, as defined in §1.2, described in sufficient technical detail to permit characterisation of its feature representation and decision procedure.	Position papers, editorials and opinion pieces; contributions describing a mechanism only at the level of a block diagram with no specification of features, parameters or decision rule.
Comparator	An explicit comparator of any kind: competing mechanisms, ablated variants of the proposed mechanism, alternative algorithms, or a formally stated hardness assumption serving as the security baseline.	Contributions reporting absolute performance with no comparator of any kind.
Outcomes	At least one quantitatively or formally reported outcome bearing on assurance (e.g. detection accuracy, bit-recovery accuracy, guessing entropy, precision, recall, advantage bound) or on the cost of assurance (e.g. runtime, communication overhead, parameter count, latency).	Contributions in which no outcome is reported, or in which outcomes are reported only qualitatively.
Study design	Peer-reviewed primary research or peer-reviewed systematic survey. Both empirical and theoretical designs eligible.	Preprints without peer review; extended abstracts; posters; theses; retracted articles.
Report characteristics	Published in the defined sampling frame (§2.3); full text available; written in English.	Outside the sampling frame; full text unavailable; not in English.

The remainder of the review is organised as follows. Section 2 sets out the methods in the detail required for independent replication, including the eligibility criteria, the search and selection procedure, the extraction form, the appraisal rubric and the synthesis strategy. Section 3 reports the results, question by question, supported by twelve tables and four figures. Section 4 discusses the principal findings, situates them against the wider literature, and articulates the cross-cutting themes. Section 5 states the limitations of the review itself, with particular candour about the bounded sampling frame. Section 6 sets out the research agenda. Section 7 concludes.

2. METHODS

The review was conducted and is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [6]. The methods were specified before extraction commenced and are reported here in full; no post hoc amendments were made to the eligibility criteria, the extraction fields or the appraisal rubric after data extraction had begun.

2.1. PROTOCOL AND REGISTRATION

A written protocol governing the research questions, eligibility criteria, extraction form, appraisal rubric and synthesis strategy was prepared in advance. The protocol was not registered in PROSPERO, since that registry accepts health-related reviews and the present topic falls outside its remit; no comparable mandatory registry exists for computer-science systematic re-

views. In lieu of registration, the protocol elements are reproduced in full in this section and in Appendices A to C, so that the review is replicable from the published record alone. Deviations from the protocol are reported in Section 5.3; one deviation occurred and is stated there.

2.2. ELIGIBILITY CRITERIA

Eligibility was specified using a PICOS-style framework adapted to computational security research, in which Population is replaced by the class of system or artefact under protection, Intervention by the assurance mechanism, Comparator by the baseline against which the mechanism is assessed, Outcomes by the reported security and cost measures, and Study design by the form of the contribution. The full criteria are given in Table 1.

2.3. INFORMATION SOURCES AND SAMPLING FRAME

The sampling frame was defined a priori, and this definition is the most consequential methodological decision in the review. Rather than attempting an exhaustive multi-database retrieval across five heterogeneous subfields (an undertaking that would have produced a corpus far too large to appraise at the item level, and one whose coverage claims would have been difficult to substantiate), the review adopts a bounded frame: the peer-reviewed research published in *Cyber Security and Applications*, Volume 4 (2026), an open-access journal published by Elsevier on behalf of KeAi Communications, as enumerated on the cutoff date of 15 September 2026. The frame is therefore the complete set of

Table 2. Information sources, enumeration parameters and yield.

Parameter	Specification
Source	Cyber Security and Applications (ISSN 2772-9184), Elsevier B.V. on behalf of KeAi Communications Co., Ltd.
Volume enumerated	Volume 4 (2026)
Access mode	Open access; full texts retrieved as publisher PDFs of record
Article-type filter	Peer-reviewed research articles and peer-reviewed surveys
Language limit	English
Supplementary methods	None. Citation chasing was excluded by design (see §2.4)
Records identified	Nine (article numbers 100115, 100124, 100125, 100126, 100127, 100128, 100129, 100130 and 100132).
Duplicates removed	0
Records screened	Nine at title/abstract; five progressed to full-text assessment.
Full texts assessed	Five.
Records included	Five (article numbers 100115, 100124, 100125, 100128 and 100129).
Grey literature	Not sought; excluded by the study-design criterion of Table 1
Enumeration cutoff date	15 September 2026
Records excluded at title/abstract screening	Four (article numbers 100126, 100127, 100130 and 100132) on Criterion I of Table 1.

Table 3. Data extraction form: fields, definitions and mapping to review questions.

#	Domain	Field	RQ
1	Bibliographic	Authors, year, title, article number, digital object identifier	–
2	Bibliographic	Institutional affiliation and country of corresponding author	–
3	Bibliographic	Declared funding sources	–
4	Bibliographic	Declared conflicts of interest	–
5	Design	Contribution type (primary experimental, theoretical, systematic survey)	–
6	Design	Assurance class (detective, constructive, evaluative)	RQ2
7	Design	Assurance layer (application, media, network, protocol, implementation)	RQ2
8	Threat	Protected asset and security goal	RQ1
9	Threat	Adversary capabilities assumed	RQ1
10	Threat	Evasion or adaptive-adversary behaviour considered	RQ1
11	Mechanism	Core technique and algorithmic family	RQ2
12	Mechanism	Feature representation and level of abstraction	RQ2
13	Mechanism	Decision procedure or verification equation	RQ2
14	Mechanism	Key hyper-parameters or scheme parameters	RQ2
15	Evaluation	Datasets, sample sizes and public availability	RQ3
16	Evaluation	Evaluation metrics and their definitions	RQ3
17	Evaluation	Baselines or comparators used	RQ4
18	Evaluation	Computational environment and implementation stack	RQ3
19	Outcomes	Primary assurance outcomes with numerical values	RQ4
20	Outcomes	Cost, overhead and complexity outcomes	RQ4
21	Outcomes	Variance, confidence intervals or repeated-run statistics	RQ4
22	Limitations	Limitations conceded by the primary authors	RQ5

Volume 4 records available at that cutoff, and any record subsequently added to Volume 4 after the cutoff is by construction outside the frame and is not considered in this review.

The rationale for this design is threefold. First, the frame is exhaustively enumerable and therefore verifiable: any reader can confirm that the records identified are precisely the records available, which is a property that broad multi-database searches rarely possess in practice. Second, the frame holds constant the peer-review standard, the editorial scope and the publication window, so that variation observed between studies is attributable to the research rather than to venue heterogeneity. Third, because

the journal's scope deliberately spans applied cryptography, network security, software security and machine learning for security, a single volume yields a cross-section of layers that a topically defined search would have fragmented.

The limitations that follow from this design are substantial and are not minimised. They are stated in full in Section 5.1, where the consequences for generalisability and for the interpretation of the term 'systematic' are addressed directly. The reader is asked to read the results of this review as a rigorous and reproducible synthesis of a defined and bounded evidence base, not as an exhaustive census of the world literature on assurance mechanisms.

Table 4. Eight-domain methodological quality appraisal rubric with anchor descriptors.

Domain	Construct assessed	Anchor for a score of 2
Q1 Objectives	Clarity and specificity of the research question and stated contribution	Question and contributions stated explicitly and enumerated; scope boundaries defined
Q2 Threat model	Completeness of the adversary model and security-goal specification	Adversary capabilities, goals and interaction model specified explicitly, formally or operationally
Q3 Transparency	Sufficiency of procedural detail for independent replication	Parameters, environment, versions and procedure reported completely and without internal inconsistency
Q4 Data / benchmark	Appropriateness, scale and availability of the evaluation data	Multiple appropriate datasets, publicly available, with sample sizes and splits stated
Q5 Metrics	Appropriateness, completeness and accurate reporting of outcome measures	Metrics defined formally, appropriate to the goal, covering both assurance and cost, reported consistently
Q6 Baselines	Adequacy and fairness of the comparison	Multiple relevant comparators, evaluated under matched conditions or by formal reduction
Q7 Analytical rigour	Strength of the inferential apparatus supporting the claims	Repeated runs with dispersion statistics, ablation or sensitivity analysis, or a complete security reduction
Q8 Limitations	Candour and specificity in acknowledging constraints	Specific, substantive limitations conceded and their consequences for the claims discussed

2.4. SEARCH STRATEGY

Because the sampling frame is an enumerable set rather than a queryable index, the 'search' consisted of complete enumeration of the frame followed by application of the eligibility criteria. For transparency and to support replication, the equivalent Boolean expression that would retrieve this evidence base from a general bibliographic index is recorded in Appendix A, together with the field limits and date limits that would be required. Enumeration was performed once, and the final date of enumeration was recorded. No supplementary snowballing was undertaken, because forward and backward citation chasing would by construction retrieve records outside the defined frame and would therefore have violated the eligibility criteria of Table 1. Table 2 documents the information sources.

2.5. SELECTION PROCESS

Selection proceeded in two stages. At the first stage, titles, abstracts and keyword lists of all enumerated records were screened against the eligibility criteria. At the second stage, the full text of every record surviving the first stage was assessed against the same criteria, with particular attention to the comparator and outcome requirements, which are the two criteria most likely to exclude an otherwise relevant contribution.

Screening and full-text assessment were performed independently and in duplicate by two reviewers working blind to one another's decisions. Each reviewer recorded, for every record, an include or exclude decision together with the specific criterion of Table 1 on which that decision turned. The two sets of decisions were then compared, and any record on which the reviewers disagreed was referred to a third reviewer whose determination was final. The complete decision record is reproduced in Appendix B.

Agreement between the two independent reviewers is reported in Appendix B, together with an explanation of why a chance-corrected agreement coefficient cannot be computed for a corpus in which every record falls into a single category. No record

required adjudication by the third reviewer, since no disagreement arose; the adjudication procedure was nevertheless specified in advance and would have been applied had disagreement occurred.

No automation tools, machine-learning classifiers or automated de-duplication utilities were used at any stage of selection. All screening was performed manually.

2.6. DATA COLLECTION PROCESS AND DATA ITEMS

A structured extraction form comprising twenty-two fields across six domains was developed before extraction and piloted on one record; no revisions to the form proved necessary after the pilot, and the pilot record was retained in the final dataset. Extraction was performed independently and in duplicate by two reviewers, working directly from the full texts including tables and figures, and the two completed forms were reconciled field by field, with any discrepancy resolved by returning to the primary source and, where necessary, by referral to a third reviewer. Where a value appeared in more than one location within a paper, the value reported in the results tables was taken as authoritative over the value reported in the abstract, and any discrepancy between the two was recorded as a reporting-consistency observation and carried forward to the quality appraisal. One such discrepancy was identified and is reported in Section 3.3.3.

No assumptions were made about missing data and no values were imputed. Where a field was not reported by a study, it is recorded as 'not reported' throughout, and where a field is inapplicable to a study's design it is recorded as 'not applicable'; the two are distinguished consistently, since conflating them would systematically penalise theoretical contributions. Table 3 lists the extraction fields.

2.7. STUDY RISK-OF-BIAS AND QUALITY ASSESSMENT

No established risk-of-bias instrument is applicable to a corpus that mixes empirical machine-learning evaluations with reduction-based cryptographic proofs and with a systematic sur-

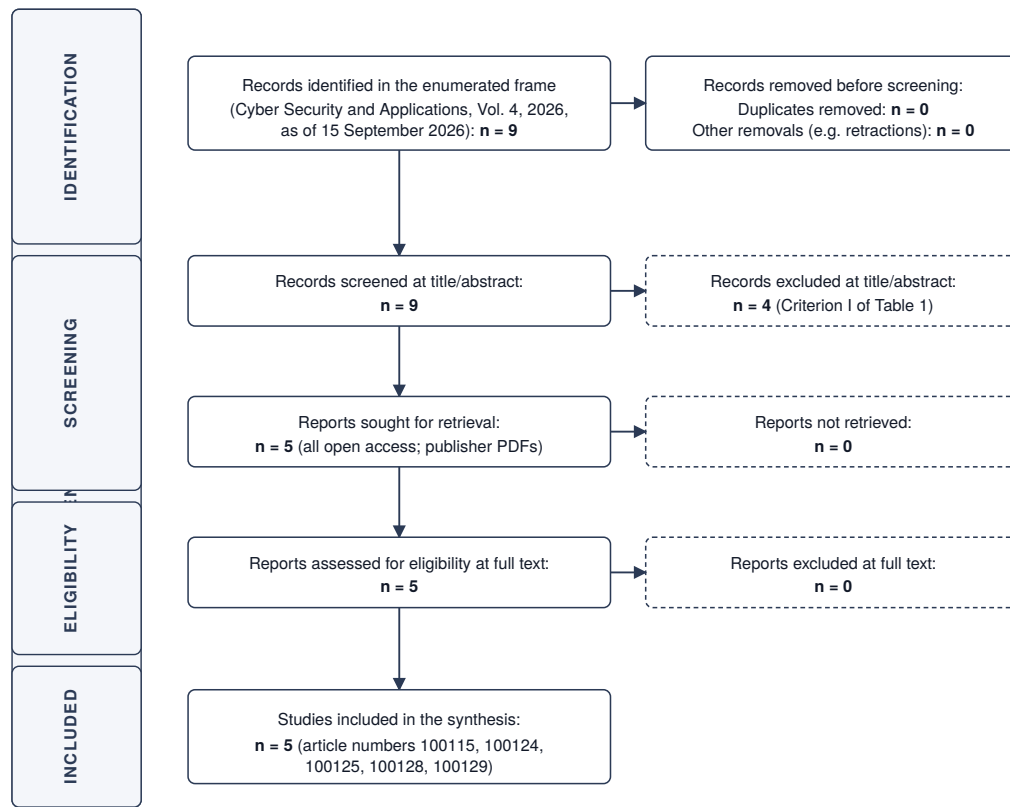


Figure 1. PRISMA 2020 flow diagram of study identification, screening, eligibility assessment and inclusion, applied to the sampling frame defined as Cyber Security and Applications, Volume 4 (2026), enumerated on 15 September 2026. Nine records were identified; four were excluded at title/abstract screening on Criterion I of Table 1 (no assurance mechanism proposed); five proceeded to full-text assessment, satisfied every inclusion criterion and were included in the synthesis.

Table 5. Bibliographic and design characteristics of the included studies.

ID	Study	Origin	Article no.	Contribution type	Assurance class	Layer
S1	Wu, Liu, Lin, Zhao, Zhou, Wu & Hong [1]	Zhejiang Univ. of Technology, Zhejiang Univ.; NSFOCUS (China)	100124	Systematic survey (63 primary studies, 2012–2025)	Detective (synthesised)	L1 Application artefact
S2	Wen, Tan, Li & Hu [2]	Hubei Minzu University (China)	100129	Primary experimental	Constructive	L2 Media object
S3	Achaal, Adda, Berger & Awde [3]	Univ. du Québec à Rimouski; CÉGEP de Sainte-Foy (Canada)	100115	Primary experimental	Detective	L3 Network / device
S4	Zhao, Liu, Zhou, Yang & Zhang [4]	Henan Women's Vocational College; Shaanxi Normal Univ.; Hubei Univ. of Technology (China)	100128	Primary theoretical with empirical validation	Constructive	L4 Cryptographic protocol
S5	Amrouche, Boubchir & Yahiaoui [5]	LIASD, Univ. of Paris 8 (France); CERIST, Algiers (Algeria)	100125	Primary experimental	Evaluative (offensive)	L5 Implementation

vey. Instruments developed for clinical research presuppose designs and outcome structures that do not obtain here. An eight-domain rubric was therefore constructed for this review, drawing on the appraisal logic common to established instruments –

namely that quality is a function of the clarity of the question, the adequacy of the design to that question, the transparency of the procedure, the appropriateness of the comparison, the rigour of the analysis and the candour of the reporting. Each domain

Table 6. Scope, objective and headline claim of each included study.

ID	Stated objective	Headline claim as advanced by the authors
S1	To review progress in repackaging detection for Android applications, evaluate the advantages and limitations of each technique family, and identify future directions.	A six-category taxonomy is established that extends prior surveys beyond 2019 to 2025, refines the taxonomy by distinguishing static from dynamic user-interface detection, and documents emerging evasion tactics including interface obfuscation.
S2	To design an audio watermarking scheme achieving simultaneous imperceptibility and robustness by improving the fusion of watermark and spectral features.	A dual-encoder U-Net with multi-scale fusion attains higher imperceptibility than comparators on three public corpora while achieving bit-recovery accuracy approaching complete recovery under most attacks.
S3	To authenticate smart meters and detect identity-based attacks in advanced metering infrastructure in real time, without predefined signal profiles or static thresholds.	A three-phase framework combining physics-based signal estimation, unsupervised local anomaly detection and cooperative classification achieves high detection sensitivity while maintaining a low false-alarm rate.
S4	To construct a multi-receiver, multi-message hybrid signcryption scheme providing receiver anonymity without an explicit recipient list.	Removing the explicit receiver list preserves confidentiality and unforgeability while achieving stronger anonymity and superior computational efficiency relative to comparator schemes.
S5	To integrate metaheuristic algorithms into neural architecture search for profiled side-channel analysis and compare them against conventional search strategies.	Metaheuristic-driven architecture search consistently outperforms random search and Bayesian optimisation, with population-based methods superior in complex search spaces and single-solution methods offering favourable time–performance trade-offs in simpler ones.

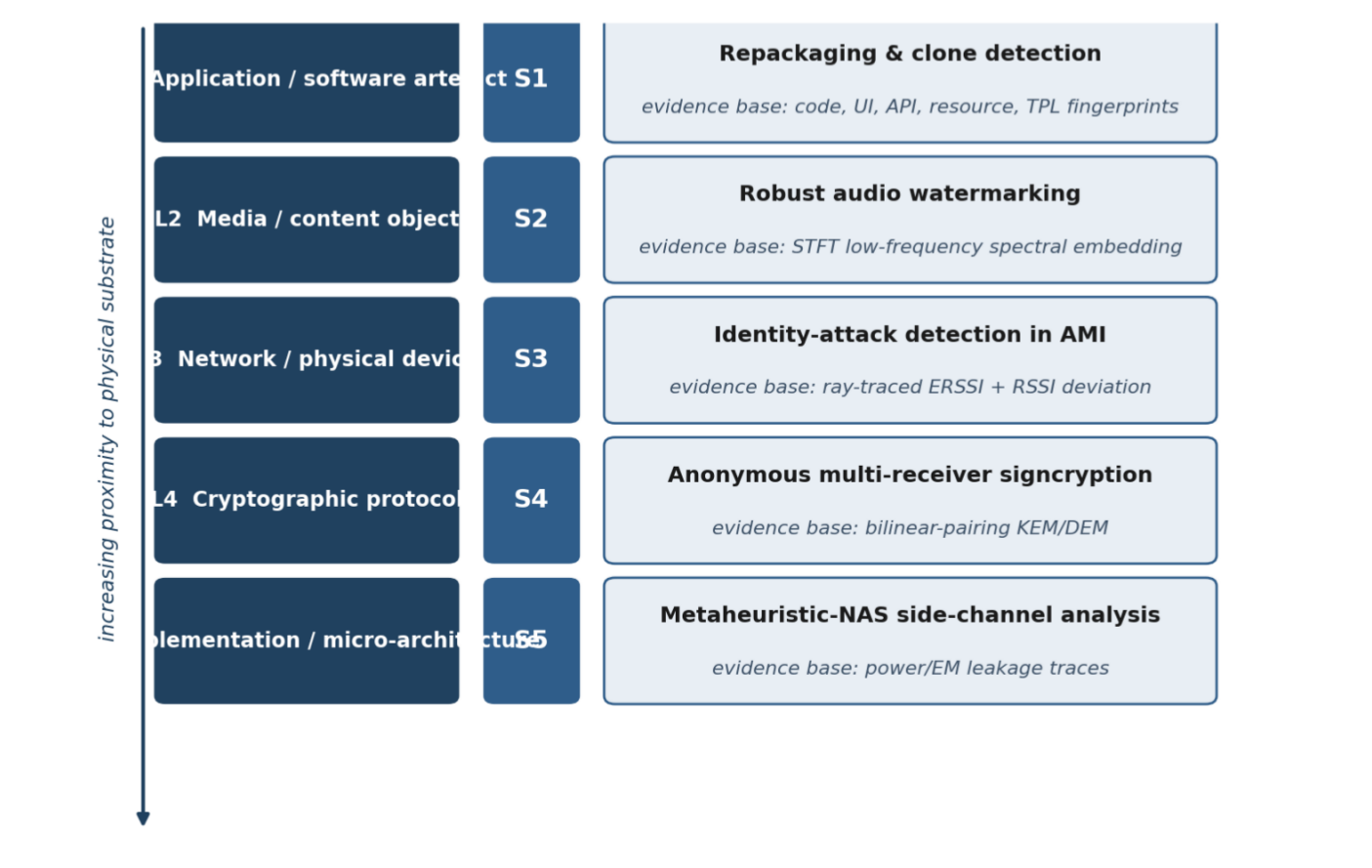


Figure 2. Cross-layer map of the assurance mechanisms represented in the included evidence base, ordered by increasing proximity to the physical substrate. Each layer is occupied by exactly one study, so the corpus provides breadth without within-layer replication.

is scored on a three-point scale: 0 where the domain is not addressed, 1 where it is partially addressed, and 2 where it is ade-

quately addressed. The maximum attainable score is 16. Where a domain is genuinely inapplicable to a study’s design,

S1 Wu et al.	2	2	2	1	1	2	1	2	81.3%
S2 Wen et al.	2	2	2	2	2	2	1	1	87.5%
S3 Achaal et al.	2	2	2	1	2	1	1	2	81.3%
S4 Zhao et al.	2	2	1	n/a	2	2	2	2	92.9%
S5 Amrouche et al.	2	2	2	2	1	2	2	2	93.8%
	Q1 Objectives	Q2 Threat model	Q3 Transparency	Q4 Data / benchmark	Q5 Metrics	Q6 Baselines	Q7 Analytical rigour	Q8 Limitations	

Figure 3. Item-level methodological quality appraisal. Scores are 0 (domain not addressed), 1 (partially addressed) or 2 (adequately addressed); 'n/a' denotes a domain inapplicable to the study's design and excluded from the denominator. Overall scores are expressed as a percentage of the attainable maximum.

it is scored 'not applicable' and excluded from both numerator and denominator, and the overall score is reported as a percentage of the attainable maximum. This is essential for fairness: a theoretical cryptographic construction has no dataset, and scoring it zero on a dataset domain would measure its genre rather than its quality. The rubric and its anchor descriptors are given in Table 4 and reproduced with full scoring rationale in Appendix C.

Quality appraisal was performed independently and in duplicate by two reviewers, and the two sets of domain scores were reconciled by discussion against the anchor descriptors of Table 4; any difference that could not be resolved by discussion was to be settled by a third reviewer, and none arose. Item-level scores are published in full in Figure 3 and Appendix C rather than reported only in aggregate, so that every judgement is exposed to external challenge and any reader who disagrees with a particular score may recompute the aggregate from the published items.

2.8. EFFECT MEASURES

No common effect measure exists across the included studies, and none was imposed. The outcome measures are constitutively different: bit-recovery accuracy under signal distortion, classification precision and recall under identity spoofing, guessing entropy as a function of the number of attack traces, and asymptotic advantage bounds expressed in a security parameter are not commensurable quantities, and any attempt to normalise them onto a single scale would create an artefact rather than a finding. Outcomes are therefore reported in their native units throughout, with the units, the conditions and the direction of improvement stated explicitly for each. Where two studies do report a genuinely comparable quantity, this is noted and the comparison is drawn.

2.9. SYNTHESIS METHODS

Meta-analysis was neither attempted nor appropriate. The precondition for statistical pooling is that the studies estimate a common underlying parameter, and that condition fails here at the most basic level. A structured narrative synthesis was therefore adopted, following a five-step procedure. First, each study was tabulated against the extraction fields to produce the evidence tables of Section 3. Second, studies were grouped by assurance layer and by assurance class to establish the structure of the evidence base. Third, within each review question, findings were compared across studies to identify convergences, divergences and gaps. Fourth, recurring patterns that were visible only across layers were abstracted into the cross-cutting themes of Section 4.3. Fifth, each identified gap was mapped to a research direction, producing the agenda of Section 6, so that every agenda item is traceable to a specific evidential deficiency.

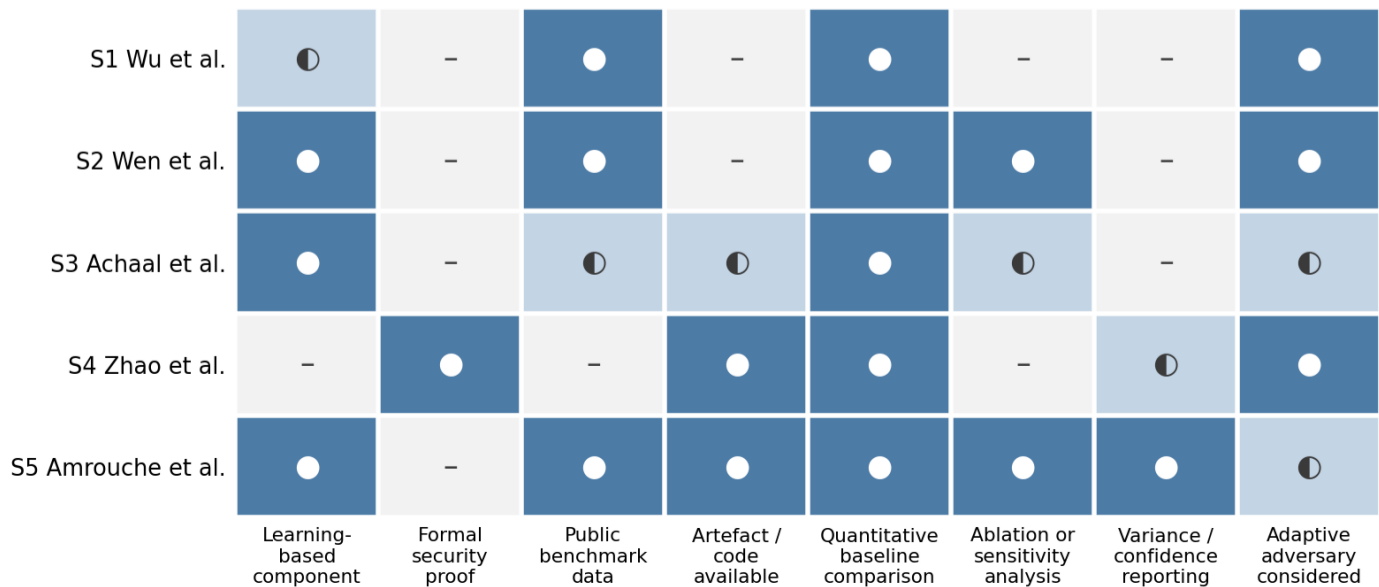
Vote counting on the basis of statistical significance was not used, since it is an unreliable synthesis method and, in any case, the included studies do not report significance tests of the required kind. Sensitivity analysis was performed by re-deriving each cross-cutting theme with each study omitted in turn, in order to establish whether any theme depended on a single study; the results of this analysis are reported in Section 4.4.

2.10. REPORTING BIAS AND CERTAINTY ASSESSMENT

Formal assessment of reporting bias across studies – funnel plots, Egger's test and related procedures – requires a body of studies estimating a common effect and is not applicable here. Two narrative checks were substituted. First, each study was examined for evidence of selective outcome reporting within the study, by comparing the outcomes promised in the abstract and introduction against the outcomes actually reported in the results. Second, the direction of reported findings was examined across the corpus for the signature of publication bias, namely a uniform preponderance of results favourable to the authors' own propos-

Table 7. Quality appraisal scores by domain, with attainable maxima.

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Total	Max	%
S1	2	2	2	1	1	2	1	2	13	16	81.3
S2	2	2	2	2	2	2	1	1	14	16	87.5
S3	2	2	2	1	2	1	1	2	13	16	81.3
S4	2	2	1	n/a	2	2	2	2	13	14	92.9
S5	2	2	2	2	1	2	2	2	15	16	93.8
Median	2	2	2	1.5	2	2	1	2	13	16	87.5

**Figure 4. Methodological and reproducibility attributes across the included studies. The distribution shows near-universal baseline comparison alongside sparse variance reporting and a single instance of formal proof.**

als.

Certainty in the body of evidence was assessed narratively against four considerations adapted from the GRADE framework: the methodological quality of the contributing studies, the directness of the evidence in relation to the review question, the consistency of findings across studies, and the precision with which effects are reported. Certainty is characterised qualitatively as high, moderate, low or very low for each principal finding, and these characterisations are reported alongside the findings in Section 4.1 rather than in a separate table, so that each claim carries its own certainty rating at the point of assertion.

3. RESULTS

3.1. STUDY SELECTION

Enumeration of the sampling frame on the cutoff date of 15 September 2026 identified nine records (Cyber Security and Applications, Volume 4, 2026, article numbers 100115, 100124, 100125, 100126, 100127, 100128, 100129, 100130 and 100132). No duplicates were present, since the frame is a single volume of a single journal. All nine records proceeded to title, abstract and keyword screening. On the basis of the criteria of Table 1, four records were excluded at title/abstract screening on grounds that they did not present an assurance mechanism in the sense of the eligibility criteria (Criterion I; no detective, constructive or evaluative mechanism for authenticity, integrity or provenance was

proposed), and the remaining five records proceeded to full-text assessment. On full-text assessment against the criteria of Table 1, all five satisfied every inclusion criterion and triggered no exclusion criterion. Five studies were therefore included in the synthesis. The selection process is depicted in Figure 1, and the record-level decisions are reproduced in full in Appendix B.

The four exclusions at title/abstract screening are recorded in Appendix B with the specific criterion invoked. Each of the four records is a research article published in Volume 4 whose topic falls outside the scope of assurance mechanisms as defined in Section 1.2: they neither detect tampering nor embed verifiable structure nor attack a deployed mechanism to quantify its assurance, and they therefore fail Criterion I of Table 1. That exclusion is a proper operation of the eligibility criteria and not evidence of frame drift; the criteria were specified before enumeration and were not relaxed. The full-text stage in turn admits every remaining record because the sampling frame is a single volume of a journal whose editorial scope is closely aligned with the review topic, so a high full-text inclusion rate is the expected consequence of frame selection rather than evidence of permissive screening. The criteria nevertheless retain discriminating power at both stages: their operation is documented in Appendix B.

Table 8. Adversary models, assumed capabilities and treatment of adaptive behaviour.

ID	Protected asset	Security goal	Adversary capabilities assumed	Adaptive / evasive behaviour
S1	Android application package and its intellectual property	Detect illegitimate derivation of one application from another	Full possession of the package; ability to decompile, modify bytecode, replace resources, load code dynamically at run time, obfuscate the user interface, and redistribute through third-party markets	Extensively treated. Interface obfuscation, packing, reflection and native-library evasion are documented, and an adversarial-example attack framework capable of mass-producing detector-evading variants is discussed
S2	Audio content object and its provenance mark	Recover the embedded mark after distortion; keep the mark imperceptible	Ability to apply signal-processing distortions: additive white Gaussian and environmental noise, fade shaping, echo, low-pass filtering, resampling, volume change, requantisation, median filtering and lossy compression, singly and in six combinations	Partially treated. Non-adaptive signal distortions are modelled comprehensively; an adversary with knowledge of the embedding network who optimises a removal attack against it is not modelled
S3	Smart-meter identity within a metering network	Detect impersonation of a legitimate meter identity	Ability to transmit using another meter's identifier from a different physical position, either while the legitimate meter is inactive or concurrently with it	Partially treated. Two impersonation scenarios are modelled; an adversary who positions itself to mimic the expected signal strength at the observer is acknowledged as a residual risk in the single-neighbour configuration
S4	Broadcast ciphertext, sender authenticity and receiver identity	Confidentiality, unforgeability and receiver anonymity	Probabilistic polynomial-time adversary with adaptive access to key-extraction, signcryption and unsigncryption oracles, subject to the standard restriction on challenge identities; permitted to declare target identities in advance	Fully treated within the formal model. Adaptive queries are granted by construction, and a guessing attack on the identity space is analysed separately with an explicit min-entropy bound
S5	Secret key of a cryptographic implementation	Recover the key from physical leakage (adversary's goal, not defender's)	Profiling adversary in possession of a device identical to the target with control of the key, able to acquire leakage traces and construct statistical models before attacking	Partially treated. The adversary is the study's own subject; countermeasure adaptation by the defender, such as masking beyond that already present in one dataset, is not varied experimentally

3.2. CHARACTERISTICS OF INCLUDED STUDIES

The five included studies are designated S1 to S5 throughout this review, in the order in which they are introduced. Their bibliographic and design characteristics are summarised in Table 5, and their substantive scope is summarised in Table 6.

The corpus is geographically concentrated but not uniform: three studies originate from institutions in China, one from Canada and one jointly from France and Algeria. Four are primary contributions and one is a systematic survey; among the primary contributions, three are principally empirical and one is principally theoretical with an empirical component. In terms of assurance class, three studies are constructive, one is detective and one is evaluative. In terms of assurance layer, each of the five occupies a distinct layer, so that the corpus contains no within-layer replication whatever – a structural feature of the evidence base with important consequences for the certainty of the findings, addressed in Section 4.1.

3.3. METHODOLOGICAL QUALITY OF THE INCLUDED STUDIES

3.3.1. Overall appraisal

Aggregate quality was moderate to high across the corpus, ranging from 81.3 per cent to 93.8 per cent of the attainable score. No study fell below the threshold at which its findings would warrant discounting, and no study attained a perfect score. Item-level results are displayed in Figure 3 and the scoring rationale for every cell is set out in Appendix C.

3.3.2. Domains of consistent strength

Two domains scored uniformly at the maximum. Every study stated its objectives explicitly and enumerated its contributions (Q1), and every study specified an adversary model with sufficient completeness to determine what the mechanism is and is not intended to resist (Q2). The character of that specification differed markedly by genre – S4 defines its adversary through three formal security games with explicitly bounded oracle access, whereas S2 defines its adversary operationally through a catalogue of ten signal-processing attacks with parameter ranges – but in both cases the reader can determine the boundary of the security claim, which is the property the domain measures.

Comparator adequacy (Q6) was also strong, scoring 2 in four

of five studies. S2 retrained all three comparator systems on the same corpus at a fixed hundred-bit payload, which is a materially stronger comparison than citing published figures obtained under different conditions. S5 compared four metaheuristics against two conventional strategies under an explicitly matched evaluation budget, holding the product of population size and generation count constant at fifty evaluations so that observed differences are attributable to search behaviour rather than to computational expenditure. S4 compared against three published schemes on identical hardware with ten repetitions. These are exemplary comparison designs.

3.3.3. Domains of consistent weakness

Analytical rigour (Q7) was the weakest domain, with a median score of 1. Only S5 reported dispersion statistics from repeated experimentation, computing mean rank-one values, standard deviations and ninety-five per cent confidence intervals over one hundred independent runs. S4 averaged ten executions but reported only means. S2 and S3 report point estimates without any measure of variability, notwithstanding that both employ stochastic training procedures whose outcomes vary across initialisations. This matters directly for interpretation: where S2 reports a mean bit-recovery accuracy of 99.76 per cent against a comparator's 98.29 per cent, the reader cannot determine whether the difference exceeds run-to-run variation, and the same limitation attaches to S3's reported margin over its comparator. The finding is not that these results are wrong, but that the evidence does not establish the precision the presentation implies.

Two studies were partially downgraded for reporting consistency. In S5, the abstract attributes the recovery of the correct key at guessing entropy of one within approximately sixty traces in the multilayer-perceptron search space to simulated annealing, whereas the corresponding results table attributes the value of sixty traces to tabu search and records a materially different value for simulated annealing under the same configuration. Following the extraction protocol of Section 2.6, the tabulated value was taken as authoritative and the discrepancy recorded, reducing the score on Q5. The substantive conclusion of the study – that metaheuristic search outperforms conventional search – is unaffected, since it rests on the tabulated results as a whole rather than on the attribution of any single value; but the discrepancy is material to a reader seeking to identify which algorithm to deploy. In S4, the parameter used in the receiver-authorisation component is defined in the scheme construction as a hash of both an ephemeral group element and the receiver identity, while the corresponding definition in the deployment protocol section omits the ephemeral element; since the study's own anonymity argument depends on the inclusion of that element to prevent the index value from being linkable across sessions, the omission is an internal inconsistency in a security-relevant definition and reduced the score on Q3.

Limitation reporting (Q8) was strong in three studies and partial in one. S3 is notably candid, stating explicitly that it was difficult to identify a dataset meeting the study's requirements and that the evaluation consequently rests on a single real-world corpus. S4 concedes plainly that sender anonymity is not achieved and that the legitimacy determination for anonymous revocation

requests lies outside its scope. S1 devotes a substantial portion of its conclusion to the limitations of the field it surveys. S2, by contrast, confines its acknowledged limitations to the observation that watermark localisation remains future work, without addressing the reliance on post-hoc embedding, the absence of adaptive-adversary evaluation, or the asymmetry in model size between encoder and decoder.

3.3.4. Reproducibility posture

Artefact availability varied considerably and is summarised, together with other methodological attributes, in Figure 4. S4 and S5 provide the strongest posture: S4 publishes an implementation repository, and S5 both uses exclusively public benchmark corpora and adopts a search space defined in prior work for which an implementation is publicly available. S3 states that code is available upon request, a position that is better than silence but that does not permit independent verification without author cooperation. S2 reports its hyper-parameters, transform settings, optimiser configuration and hardware in sufficient detail to support reimplementing, but makes no artefact-availability statement. S1, as a survey, has no implementation artefact, but documents its retrieval protocol in sufficient detail – databases, Boolean expression, retrieval window, yield at each filtering stage – for the search to be repeated, which is the corresponding reproducibility obligation for its genre and one that many surveys do not discharge.

3.4. RQ1: ADVERSARY MODELS AND THREAT CAPABILITIES

Adversary specification was, as noted, a domain of uniform strength, but the models differ profoundly in structure, in the powers granted to the adversary and in whether adaptation is contemplated. Table 8 sets out the extracted threat models.

Three observations follow. First, the profiling adversary of S5 is the strongest adversary in the corpus by a considerable margin, since it is granted an identical device with a known key – a capability that would be regarded as extravagant at any other layer. This is not a flaw but a deliberate methodological convention in implementation security, where the purpose is to establish worst-case leakage. The consequence for cross-layer reasoning is nevertheless significant: assurance claims at the implementation layer are calibrated against a far more powerful adversary than assurance claims at the media or network layers, so numerically comparable success rates at different layers do not represent comparable security.

Second, only S4 grants its adversary adaptive oracle access as a structural feature of the model. The learning-based studies model adversaries who apply transformations drawn from a fixed catalogue, and while S1 documents adaptive adversaries extensively in the literature it surveys, that documentation appears as a threat characterisation rather than as an evaluation condition to which the surveyed mechanisms have been subjected. The gap between the adversary that is described and the adversary against which mechanisms are actually tested is, on this evidence, widest precisely where learning-based methods are most heavily used.

Third, S1 supplies what the other studies lack: a documented instance of the adaptive threat materialising. The survey records an attack framework that combines the insertion of malicious code into a benign carrier with adversarial-example generation,

Table 9. Technical mechanisms, feature representations and decision procedures.

ID	Core technique	Feature representation	Decision procedure	Abstract level
S1	Six-family taxonomy: code fingerprinting; user-interface analysis; application-programming-interface analysis; resource-file analysis; third-party-library identification; alternative approaches including watermarking, logic bombs, traffic analysis and authorship attribution	Structural fingerprints (call-graph topology, control-flow structure); token-level fingerprints (opcode n-grams, source tokens); semantic fingerprints (input–output state behaviour); interface layout trees and region graphs; call-dependency graphs; resource hashes and icon embeddings	Similarity thresholding over the chosen representation, using set, vector or graph similarity; increasingly, supervised or graph-neural classification	Ranges from token to semantic; the taxonomy is explicitly ordered by abstraction
S2	Dual-encoder U-Net with multi-scale feature fusion, operating on the short-time Fourier transform, trained end-to-end through a learned distortion simulator	Magnitude spectrogram partitioned at eight kilohertz into low- and high-frequency components; the watermark is expanded along the temporal axis and fused with spectral features by channel-wise concatenation, then processed through parallel convolutional paths of three distinct receptive-field sizes	Local maximum-pooling over each time segment recovers the bit sequence; training minimises a weighted sum of spectral reconstruction error and bit-recovery error at weights of one hundred and one respectively	Signal-domain, at a deliberately chosen frequency band
S3	Three-phase pipeline: ray-traced propagation modelling corrected by random-forest regression; unsupervised isolation-forest anomaly detection; cooperative random-forest classification among neighbours	Expected received signal strength derived from geometry, materials and weather; the absolute deviation between measured and expected strength together with inter-node distance; and, at the cooperative stage, cosine similarity across all neighbours plus the mean and standard deviation of local deviation	Isolation depth aggregated across the forest yields an anomaly score in place of a static threshold; a majority vote across the cooperative forest confirms or rejects the local alert	Physical, grounded in electromagnetic propagation
S4	Identity-based hybrid signcryption in the key- and data-encapsulation paradigm, over bilinear pairings on an elliptic curve	Algebraic: private keys are triples derived from the master secret and per-user randomness; the recipient set is represented implicitly by a vector of pairing-derived index values rather than by an explicit identity list	Verification of a pairing equation establishes sender authenticity; membership of the computed index in the transmitted vector establishes authorisation; the session key is recovered by a pairing computation and hashed to key the symmetric layer	Algebraic, with security reduced to computational assumptions
S5	Metaheuristic neural architecture search: particle swarm optimisation and genetic algorithm (population-based); simulated annealing and tabu search (single-solution); compared against random search and Bayesian optimisation	Candidate architectures encoded as vectors over a joint space of architectural and training parameters, including layer counts, neuron and filter counts, kernel sizes, pooling configuration, activation functions and learning rate	Candidates are trained at low fidelity, scored by an attack-relevant fitness function, and the search state is updated by the metaheuristic's own rule; the selected architecture is then retrained to convergence	Meta-level: the object optimised is the model that consumes physical leakage

reporting that tailored perturbations inserted into the malicious segment allow high-volume generation of variants that evade

both state-of-the-art academic detectors and commercial engines at high success rates [7]. This is the single most consequential threat datum in the corpus, because it establishes that the evasion problem is not hypothetical, and because it applies in principle to every learning-based mechanism represented here, not only to the layer at which it was demonstrated.

3.5. RQ2: MECHANISMS AND FEATURE REPRESENTATIONS

Table 9 sets out the mechanisms, their feature representations and their decision procedures. The organising variable that emerges is the level of abstraction at which each mechanism forms its representation.

The clearest technical result concerning abstraction comes from S1, which reconciles a long-standing ambiguity in its field by separating structural, token-level and semantic fingerprints into distinct categories and assigning each line of prior work to one of them. Related application-layer work covers repackaging benchmarks and comparisons, marketplace detection, Andro-Zoo, evasive-malware detection, third-party-library detection, multi-scale dependency fusion and logic-bomb-based detection [8–15]. The argument advanced is that token-level representations characterise similarity through instruction-sequence distributions and are therefore vulnerable to reordering and rewriting; structural representations model call-graph topology and function-dependency layout and therefore survive opcode-level perturbation in non-critical methods; and semantic representations, anchored in input–output behaviour, survive even substantial token modification provided the underlying behaviour is preserved, thereby detecting the logic-consistent but token-modified case at which the lower levels fail.

This same principle is visible, expressed in different vocabularies, in three of the remaining four studies. S2 embeds below eight kilohertz on the stated ground that low-frequency components yield stronger robustness, which is the signal-domain instance of choosing a representation stable under expected transformation, consistent with wider work on robust audio watermarking, re-recording resilience, voice-cloning detection and the development of digital audio watermarking [16–19]; the ablation reported in the study shows that its multi-scale fusion modules improve average accuracy under attack from 99.71 to 99.94 per cent while reducing the distortion-free signal-to-noise ratio from 34.48 to 32.41 decibels, a direct and quantified exhibition of the robustness–imperceptibility trade-off that the choice of representation governs. S3 grounds its representation in ray-traced propagation physics precisely because a physically derived expectation, unlike a historical fingerprint, remains valid under environmental change and requires no profiling period. The method is grounded in established work on isolation forests, physical-layer authentication, radio-propagation ray tracing and smart-grid cybersecurity [20–23]. S4 – a multi-receiver anonymous identity-based hybrid signcryption scheme, hereafter designated MRA-IBHS – achieves the strongest form of the property by construction: its representation is algebraic, and the security of the representation is reduced to computational hardness rather than established empirically. The construction belongs to the signcryption, hybrid-encryption and multi-receiver anonymity line of work [24–28].

S5 stands outside this pattern and clarifies it by contrast. Its

object is not a representation of the protected artefact but a search over representations, and its finding is that the appropriate search strategy is itself a function of the structure of the space searched. This evaluation sits within prior work on automated tuning, deep-learning-based side-channel analysis, classical leakage attacks, neural architecture search and benchmark attack methodology [29–36]. Population-based methods, which maintain diversity across many candidate solutions simultaneously, are reported as superior in the larger convolutional space, whereas single-solution methods, which exploit locally, perform competitively in the smaller perceptron space at substantially lower cost – simulated annealing completing its search in 715.84 seconds against 1 952.91 seconds for random search under one configuration. This is a structural finding about optimisation landscapes rather than about assurance as such, but it bears directly on assurance because it determines how efficiently the strength of an implementation can be evaluated.

3.6. RQ3: DATASETS, BENCHMARKS AND EVALUATION METRICS

3.6.1. Datasets and benchmark maturity

Benchmark maturity is markedly uneven across the layers, and Table 10 sets out the position. Three layers are served by established public corpora; two are not. The media-layer evaluation uses the Free Music Archive and LibriSpeech corpora [37, 38], while the network-layer evaluation is based on the outdoor LoRa connectivity corpus [39].

Two structural deficiencies emerge. At the network layer, S3 could not locate a corpus containing the conjunction of properties its method requires, and therefore constructed its evaluation by augmenting a corpus collected for an unrelated purpose and injecting simulated attacks. The construction is described transparently and is defensible – genuine signal-strength measurements are preserved and only identifiers are reassigned, so the physical realism of the traces is retained – but the attacks are synthetic, and no independent corpus exists against which the resulting detection rates could be corroborated.

At the protocol layer the deficiency is of a different kind. There is no benchmark corpus for signcryption schemes and, given the nature of the artefact, there could not straightforwardly be one; the analogue of a benchmark is a standard implementation environment and a standard operation-cost model. S4 approximates this by reimplementing three comparator schemes on identical hardware and by reporting costs in a common notation of elliptic-curve and pairing operations, which is the strongest available substitute. The absence of a community-standard cost model nevertheless means that comparative efficiency claims in this area rest on the fidelity of each authors' own reimplementation of their competitors' schemes.

3.6.2. Evaluation metrics

Metric usage is heterogeneous in form but convergent in structure: every study reports at least one assurance-side measure and, with the partial exception of S1, at least one cost-side measure. Table 11 records the metrics and their operational definitions.

The most sophisticated metric practice in the corpus belongs to S5, which reports not only attack-effectiveness measures but also their dispersion across one hundred runs, and which additionally

Table 10. Datasets and evaluation corpora employed or documented.

ID	Corpora	Scale and composition	Public	Benchmark maturity
S1	RePack; AndroZoo; F-Droid; Drebin; VirusShare; Malware Genome; CloneCorp; Contagio; vendor marketplaces	RePack comprises 15 297 repackaged instances covering 2 776 original applications, filtered from AndroZoo by distinct developer signature and code similarity exceeding eighty per cent. Studies surveyed report evaluation corpora from 8 applications to 1 300 000	Mixed; RePack, AndroZoo and F-Droid public	Emerging. A community benchmark exists and is widely adopted, but the survey concludes that no widely accepted standardised corpus governs the field and that scale, distribution and annotation protocols remain inconsistent
S2	Free Music Archive; GTZAN; LibriSpeech	19 753 training, 3 608 validation and 200 test clips from the first corpus; 200 test clips from each of the other two, retained as held-out generalisation sets. All resampled to forty-eight kilohertz, ten seconds per clip	Yes, all three	Mature. Established public corpora spanning music and speech, permitting a cross-domain generalisation test
S3	Outdoor long-range wireless connectivity corpus of Tian <i>et al.</i> , augmented with weather descriptors and with simulated impersonation	Collected over five months during 2021 in Shanghai; contains transmitter and receiver identifiers, signal-strength values and environmental parameters including temperature, wind speed, pressure and humidity	Underlying corpus public; augmented and attack-labelled version not released	Immature. The authors state explicitly that identifying a corpus meeting the study's requirements was difficult, the requirement being bidirectional signal-strength measurements with synchronised environmental data
S4	None	Not applicable. Security is established by reduction; performance is measured by execution timing on two hardware platforms with ten receivers and ten repetitions	Not applicable	Absent. No community benchmark exists for comparative evaluation of signcryption schemes; comparison rests on reimplementations under matched conditions
S5	ASCAD fixed-key; CHES capture-the-flag 2018 (ASCAD random-key documented but not used)	ASCAD fixed-key provides 50 000 profiling and 10 000 attack traces of 700 features each, from an eight-bit microcontroller running a first-order masked implementation, captured electromagnetically; the capture-the-flag corpus derives from an unprotected implementation on different hardware with noisier traces	Yes, both	Mature. Community benchmarks with an established evaluation convention; the study exploits the difference between them as a robustness test

reports the trainable parameter count of each selected architecture in order to test whether performance gains are attributable to increased capacity rather than to better architectural choices. The study reports that higher attack performance does not systematically correspond to larger models, and that the population-based methods frequently select compact or moderately sized architectures achieving competitive or superior results. This is a form of confound control that is unusual in the architecture-search literature and that materially strengthens the study's central claim.

3.7. RQ4: DEMONSTRATED PERFORMANCE, COST AND OVERHEAD

Outcomes are reported here in native units, grouped by study. Table 12 gives the assurance-side outcomes and Table 13 the cost-side outcomes.

Three outcomes merit particular emphasis because they are

unusually well supported by the design that produced them.

The generalisation result in S2 is strong because of where it is strongest. The proposed method attains its highest mean accuracy, 99.91 per cent, on the speech corpus, which is the corpus furthest in character from the music on which the model was trained, and its margin over the nearest comparator is widest there: 99.91 against 94.03 per cent, a gap of nearly six percentage points, whereas the corresponding gap on the training-domain corpus is under one and a half points. A method that degrades least where domain shift is greatest is exhibiting genuine generalisation rather than corpus-specific fitting. The counterpart observation is that the perceptual cost also rises with domain shift, the signal-to-noise ratio falling from 32.41 decibels on music to 26.11 on speech, so the robustness is partly purchased with a larger perceptual footprint on the out-of-domain material.

The architectural result in S3 is strong because the study quan-

Table 11. Evaluation metrics, definitions and direction of improvement.

ID	Assurance-side metrics	Cost-side metrics	Notes on definition
S1	Similarity coefficients as surveyed: Jaccard for set-valued features; cosine for vector-valued features; Euclidean, Manhattan and Mahalanobis distances and normalised compression distance for structural comparison	Analysis time per application pair, as reported by individual surveyed systems	The survey characterises the applicability of each family – set measures are robust to sparsity and cheap; vector measures capture semantic proximity but depend on feature weighting; distance measures suit clustering and large-scale ranking – but does not itself apply them
S2	Bit-recovery accuracy, defined as the proportion of correctly recovered watermark bits; area under the receiver-operating characteristic curve across payload sizes	Signal-to-noise ratio between original and watermarked audio, in decibels; encoder and decoder parameter counts	Signal-to-noise ratio functions here as a cost measure, quantifying the perceptual price paid for robustness; the two are explicitly traded against one another in the ablation
S3	For the estimation phase, mean absolute error, root mean square error and coefficient of determination; for the detection phases, accuracy, precision and recall, aggregated across timestamp-grouped monitoring events	Communication overhead as a proportion of baseline network payload; worst-case verification latency and expected added end-to-end delay, both analytically derived	Positives and negatives are defined at the level of the monitored node across all neighbours: a malicious node correctly flagged by at least one neighbour counts as a true positive, and a legitimate node flagged by at least one neighbour as a false positive
S4	Advantage bounds in the security games: indistinguishability under chosen-ciphertext attack; existential unforgeability under chosen-message attack; residual min-entropy of the identity space under guessing attack; automated protocol verification outcomes	Transmission cost as a function of receiver count in group and field element sizes; computation cost decomposed into point additions, scalar multiplications, exponentiations and pairings; wall-clock timings on desktop and embedded hardware	Security is asymptotic and conditional on stated hardness assumptions in the random-oracle model, and is therefore categorically different in kind from an empirical accuracy figure
S5	Key rank; guessing entropy, being the average rank of the correct key over independent experiments; success rate; leakage difference distribution and its Hamming-weight variant, with a correlation-based derivative used as a search objective	Search execution time in seconds; trainable parameter count of the selected architecture	Guessing entropy is reported as the number of attack traces required to reach a rank of one, so lower is better; a value of 'none' denotes failure to reach rank one within the trace budget

tifies the value of each phase separately rather than reporting only the assembled system. Local anomaly detection achieves recall of 0.9560 at precision of only 0.7500 – a deliberate configuration, since the contamination parameter was set to flag generously – and cooperative verification then raises precision to 0.9380 while preserving recall at 0.9520. The design converts a sensitive but imprecise local detector into a system that is both sensitive and precise, and it does so at a cost the authors bound analytically at roughly five per cent of network payload. The degraded single-neighbour configuration, in which recall falls to 0.8265, both isolates the source of the gain and produces a like-for-like comparison against an external scheme operating under the same single-observation assumption.

The efficiency result in S4 is strong because it is asymptotic rather than merely empirical. The reported reduction from quadratic to linear growth in transmission cost with respect to receiver count is a property of the scheme's structure, not of the platform on which it was measured, and it is corroborated by

timings on two hardware platforms of very different capability. The candour with which the authors report that the scheme is slightly less efficient for single-message transmission, and that on the embedded platform it is second best overall rather than best, strengthens rather than weakens confidence in the remainder of the comparison.

3.8. RQ5: LIMITATIONS CONCEDED BY THE PRIMARY AUTHORS

Table 14 records the limitations the primary authors themselves acknowledge. These are reported separately from the limitations identified by this review during appraisal, which are given in Section 3.3, so that the two are not conflated.

Read across the corpus, the conceded limitations cluster into four recurring categories. Evaluation-scope limitations are conceded by three studies: the single-corpus dependency of S3, the corpus-specific inapplicability of one leakage model in S5, and the dataset-standardisation deficit documented at length by S1.

Table 12. Principal assurance outcomes as reported by the included studies.

ID	Outcome	Value	Comparator values	Condition
S2	Mean bit-recovery accuracy across ten attack families, Free Music Archive corpus	99.76%	98.43% / 90.39% / 98.29%	Hundred-bit payload; all comparators retrained on the same corpus
S2	Mean bit-recovery accuracy, GTZAN corpus	99.77%	98.78% / 88.67% / 99.11%	Held-out generalisation corpus
S2	Mean bit-recovery accuracy, LibriSpeech corpus	99.91%	98.66% / 88.52% / 94.03%	Held-out generalisation corpus; speech rather than music
S2	Worst-case accuracy under six combined attacks, GTZAN	95.30–99.98%	Lowest comparator 75.17%	Combinations of filtering, noise, fading, compression and echo
S2	Area under the receiver-operating characteristic curve across payloads	Exceeds 96% at all payloads from 100 to 800 bits; reaches 1.000 at 100 bits under several attacks	Not reported for comparators	Payload sweep at 100, 200, 400, 600 and 800 bits
S3	Signal-strength estimation, random forest	MAE 3.7601; RMSE 4.8978; coefficient of determination 0.8087	Linear regression 8.0777 / 9.9565 / 0.3697; multilayer perceptron 6.2237 / 8.1170 / 0.4482	Ray-tracing output corrected by regression on spatial and weather features
S3	Local anomaly detection, isolation forest	Accuracy 0.8950; precision 0.7500; recall 0.9560	Support vector machine 0.7842 / 0.7000 / 0.8500	Contamination parameter 0.20; one hundred isolation trees; recall deliberately prioritised over precision
S3	Cooperative classification, random forest	Accuracy 0.9418; precision 0.9380; recall 0.9520	Multilayer perceptron 0.9218 / 0.9050 / 0.9300; logistic regression 0.9118 / 0.8900 / 0.9150	Third phase, triggered only for nodes flagged locally
S3	Full framework against an external physical-layer authentication scheme	0.9418 / 0.9380 / 0.9520	Single-neighbour configuration 0.8550 / 0.8561 / 0.8265; external comparator 0.8100 / 0.7847 / 0.8142	The single-neighbour configuration is a deliberately degraded worst case matching the comparator's operating assumptions
S4	Confidentiality	Indistinguishability under chosen-ciphertext attack; adversarial advantage reduces to the decisional bilinear Diffie–Hellman problem with advantage at least equal to the adversary's	Not applicable; asymptotic guarantee	Random-oracle model; selective security with challenge identities declared in advance
S4	Unforgeability	Existential unforgeability under adaptive chosen-message attack; reduces to the computational Diffie–Hellman problem with advantage at least the adversary's advantage divided by the base of natural logarithms	Not applicable	Random-oracle model

Scope-of-guarantee limitations are conceded by two: the absence of sender anonymity in S4 and the absence of watermark localisation in S2. Adversary-coverage limitations are conceded prin-

cially by S1, which is the only study to state directly that the mechanisms in its field lack effective defences against highly obfuscated and polymorphic transformation. Cost limitations are

Table 12. Principal assurance outcomes as reported by the included studies (continued).

ID	Outcome	Value	Comparator values	Condition
S4	Anonymity	Residual min-entropy of the identity space after polynomially many guesses is the identity length less a logarithmic term	Not applicable	Randomised index derivation prevents cross-session linkage of identical identities
S4	Formal verification	Secrecy, correspondence and integrity queries all discharged	Not applicable	Automated symbolic protocol verifier
S5	Best key recovery, perceptron search space, identity leakage model	Guessing entropy of one at approximately 60 attack traces	Random search 397; Bayesian optimisation 178; genetic algorithm 155; particle swarm 180 traces at the same configuration	ASCAD fixed-key corpus; fifty training epochs; leakage-correlation objective
S5	Best key recovery, convolutional space, identity leakage model	Guessing entropy of one at 202–290 traces	Random search failed to reach rank one at either epoch budget	Genetic algorithm and particle swarm optimisation the only consistently successful strategies
S5	Best key recovery, convolutional space, Hamming-weight leakage model	Guessing entropy of one at 419 traces	Random search 615; Bayesian optimisation failed to converge	Accuracy objective; fifty epochs
S5	Best key recovery, capture-the-flag corpus, convolutional space	273 traces	Bayesian optimisation 417; random search 894; tabu search 2 055	Ten-epoch budget; genetic algorithm best performing

conceded by S2 and S3, in the form of decoder size and density-dependent overhead respectively.

What no study concedes is instructive. None of the four studies deploying learning-based components acknowledges that its evaluation excludes an adaptive adversary optimising specifically against the deployed model, even though S1 documents in its own review of the literature that such adversaries exist and succeed. The gap between what the corpus knows collectively and what any individual study acknowledges about its own evaluation is, on this evidence, the most significant unacknowledged limitation in the body of work.

4. DISCUSSION

4.1. SUMMARY OF PRINCIPAL FINDINGS

Five principal findings emerge from the synthesis. Each is stated below with its certainty characterisation, assessed against the four considerations set out in Section 2.10.

Finding 1. Assurance mechanisms across all five layers converge on a common design principle, namely the search for a representation that is invariant under benign or expected transformation while remaining sensitive to adversarial modification, and the level of abstraction at which that representation is formed is the dominant determinant of robustness. Certainty: moderate. The finding is directly supported by S1's explicit three-level fingerprint taxonomy and by the design rationales of S2, S3 and S4, and it is internally consistent across the corpus; it is downgraded from high because each layer is represented by a single study, so the pattern rests on breadth rather than replication.

Finding 2. The corpus exhibits a pronounced asymmetry be-

tween provable and empirical assurance. Only S4 discharges its security claims by reduction to stated hardness assumptions and corroborates them by automated formal verification; the four remaining studies establish their claims by benchmark performance alone. Certainty: high. The finding is a direct and unambiguous property of the extracted data, requires no inference, and is consistent across every study in the corpus.

Finding 3. Evaluation infrastructure is mature at the media and implementation layers, emerging at the application layer, and absent at the network and protocol layers. Certainty: moderate to high. The finding is directly evidenced by the extracted dataset characteristics and, importantly, is corroborated by the primary authors themselves: S3 states that a suitable corpus could not be found, and S1 states that no widely accepted standardised corpus governs its field.

Finding 4. Adaptive adversaries are documented in the corpus but are almost entirely absent from its evaluations. S1 records an attack framework that mass-produces variants evading both academic and commercial detectors, yet none of the four studies deploying learning-based components evaluates against an adversary optimising specifically against the deployed model. Certainty: moderate. The documentation of the threat is direct and the absence from evaluation is verifiable, but the magnitude of the resulting overstatement of security cannot be estimated from this corpus.

Finding 5. Where cost is quantified, the price of assurance is consistently modest in relation to the assurance obtained, but the quantification itself is inconsistent. S3 bounds its overhead analytically at approximately 5.4 per cent of network payload;

Table 13. Principal cost, overhead and complexity outcomes.

ID	Cost dimension	Reported value
S2	Model complexity	Encoder 0.48 million parameters; decoder 2.02 million; approximately 2.5 million in total. Exceeds two comparators and is comparable to the third; the asymmetry is attributed to the multi-scale fusion module in the decoder
S2	Perceptual cost	Signal-to-noise ratio 32.41 decibels on the training-domain corpus, 28.79 and 26.11 decibels on the two held-out corpora, against comparator ranges of 21.26 to 33.20 decibels
S2	Robustness–imperceptibility trade-off	Removing the decoder fusion module raises distortion-free signal-to-noise ratio from 32.41 to 34.48 decibels while reducing mean accuracy under attack from 99.94 to 99.71 per cent
S3	Communication overhead	Cooperative verification adds approximately 5.4 per cent to baseline network payload under the stated parameters: one thousand meters, neighbour density fifteen, mean hop distance eight, one quarter of nodes flagged, fourteen participating neighbours
S3	Latency	Worst-case serialised verification delay 0.5432 seconds at ten kilobits per second; expected added end-to-end delay approximately 1.0864 seconds. The authors note that broadcasting rather than serialising the trigger reduces this substantially
S3	Architectural placement of cost	Only the third phase generates inter-meter traffic; propagation estimation is centralised and periodic, and local anomaly detection is computed on-device without network load
S4	Transmission cost	Grows as a linear function of receiver count for both single- and multiple-message transmission, against quadratic growth in the three comparators. At fifty receivers the reported totals are 56 kilobits for the proposed scheme against 1 208 kilobits for two comparators and 4 440 kilobits for the third
S4	Computation cost, desktop	Signcryption 80.0 milliseconds; unsigncryption 8.7 milliseconds; total across ten executions 887.0 milliseconds, against comparator totals of 1 345.0, 2 951.0 and 1 504.0 milliseconds
S4	Computation cost, embedded platform	Signcryption 171.3 milliseconds; unsigncryption 250.1 milliseconds; total 695.0 milliseconds, against comparator totals of 677.7, 919.0 and 766.9 milliseconds – the proposed scheme is second best overall and best in the decryption phase
S4	Efficiency profile	Slightly less efficient than comparators for single-message transmission; the advantage is realised in the multiple-message regime, where the growth rate of overhead is materially lower
S5	Search cost, perceptron space	Simulated annealing 715.84 seconds; random search 1 952.91 seconds; Bayesian optimisation 1 983.69 seconds; particle swarm 2 266.68 seconds; tabu search 5 188.17 seconds
S5	Search cost, convolutional space	Genetic algorithm 1 366.92 seconds, the lowest recorded in that space, while simultaneously achieving the best guessing entropy under the same configuration
S5	Capacity confound	Trainable parameter counts of selected architectures are reported explicitly; higher attack performance is found not to correspond systematically to larger models

S4 reduces asymptotic transmission cost from quadratic to linear in receiver count; S2 accepts a reduction of roughly two decibels in signal-to-noise ratio for a reduction of about four fifths in residual bit error under attack. Certainty: low to moderate. The individual measurements are sound but the cost dimensions are not commensurable, and two studies report no dispersion around their estimates.

Table 15 presents these findings in summary form, with the contributing studies, the nature of the supporting evidence and the certainty rating for each.

4.2. INTERPRETATION IN THE CONTEXT OF THE WIDER LITERATURE

4.2.1. The invariance principle and its lineage

The invariance principle articulated in Finding 1 is not itself new; it is implicit in the software-birthmark literature and in the classical requirement that a watermark survive the transformations a signal is expected to undergo. What the present corpus contributes is a demonstration of its scope. That the same principle organises the choice of an opcode-versus-call-graph representation in software analysis, of a frequency band in audio wa-

termarking, of a physics-derived rather than history-derived expectation in wireless authentication, and of an algebraic rather than empirical representation in cryptography, suggests that it is a structural feature of the assurance problem rather than a convention of any one field.

The principle also predicts the direction of the arms race, and S1's evidence is consistent with that prediction. As attackers acquired the ability to perturb token sequences cheaply, detection research moved upward in abstraction to structural and then to semantic representations; the survey's account of this progression, culminating in behavioural fingerprints anchored in input–output state, is a documented instance of the predicted movement. The same logic implies that mechanisms whose representations cannot be raised further in abstraction – because they are already anchored in physics or in algebra – are structurally more durable, which is an argument in favour of the approaches taken by S3 and S4 that neither study makes explicitly on its own behalf.

4.2.2. Provable and empirical assurance

The asymmetry recorded in Finding 2 reflects a genuine methodological divide with deep roots. Provable security in the tradition

Table 14. Limitations conceded by the authors of the included studies.

ID	Conceded limitation
S1	The field is described as having been overshadowed by more fashionable directions in mobile security, leaving some techniques confined to traditional static-analysis and feature-comparison paradigms with limited responsiveness to emerging attacks. Datasets are characterised as limited in scale, uneven in sample distribution and lacking standardised annotation protocols, which restricts fair comparison. Most methods are said to rely on predefined rules or fixed features, lacking effective mechanisms against highly obfuscated and polymorphic repackaging. Third-party library detection is identified as accurate at library level but weak at version level, which is described as a critical bottleneck. The survey's own frequency figures are non-mutually-exclusive by construction, so category counts exceed the total number of reviewed papers.
S2	Watermark localisation is not implemented and is deferred to future work. The scheme remains a post-processing method rather than one integrated with synthesis. The decoder is substantially larger than the encoder. The authors additionally disclose the use of a large language model for English-language editing, with responsibility for content retained.
S3	It was difficult to identify a dataset meeting the study's requirements – namely bidirectional signal-strength measurements with synchronised environmental data – and the evaluation therefore rests on a single real-world corpus. Generalisation across urban, suburban and rural deployment contexts is explicitly identified as requiring further testing and is named as the principal direction for future work. Cooperative-phase overhead scales with neighbour density, and the authors propose capping the number of participating neighbours to bound it. Code is available only on request.
S4	Sender anonymity is not achieved; only receiver anonymity is provided, and the authors state that the sender's anonymity is not adequately ensured in real-world environments. Determining the legitimacy of anonymous revocation requests is outside the scope of the present work and is deferred. Security is proved in the random-oracle model rather than the standard model. The identity space must be represented in full, which may require substantial local storage at the sender.
S5	The identity leakage model could not be applied to the second corpus, because training and validation sets use different fixed keys while identity-model intermediate values depend directly on the key used during training. Convolutional accuracy results on that corpus were consistently poor and are not reported. Parameter sensitivity was investigated for one metaheuristic only. Model parameter count is not treated as an explicit optimisation objective, which the authors identify as an interesting extension. Existing architecture-search frameworks from other domains rely on spatial or graph-structured representations that do not transfer directly to one-dimensional leakage traces.

descending from the signcryption primitive [24] and the key- and data-encapsulation framework [25, 26] asks what an adversary cannot do given stated assumptions; empirical evaluation asks what a system does on data. The former yields conditional but general guarantees; the latter yields unconditional but particular ones. Neither is dispensable, and the difficulty is not that both exist but that claims arising from them are frequently juxtaposed as if commensurable.

The consequence for practice is concrete. A composite system employing the MRA-IBHS scheme of S4 to protect the reports generated by the metering framework of S3 would carry a guarantee of a very different character at each layer: at the protocol layer, unforgeability conditional on the computational Diffie–Hellman assumption holding in the chosen group; at the network layer, a detection rate of 94.18 per cent on one corpus with attacks the evaluators themselves injected. The composite is no stronger than its weakest guarantee, and identifying which guarantee is weakest requires comparing objects of fundamentally different kinds. The corpus provides no methodology for that comparison, and the wider literature, so far as the surveyed material indicates, does not supply one either.

4.2.3. The double-edged role of machine learning

Machine learning appears in four of the five studies, and the corpus illustrates all three of its roles in this domain simultaneously. It serves as the assurance mechanism in S2 and S3; as the object of study in the detection systems S1 surveys; and as the in-

strument of attack in S5, where architecture search is applied to make key recovery more effective. Its role in defeating assurance is documented most sharply by the attack framework S1 records, which merges code injection with adversarial-example generation to evade detectors at scale [7], a result that instantiates in the mobile-security setting the problem-space adversarial attack analysed in the wider literature [40].

The implication is that in adversarial settings the deployment of learning-based assurance is not a purely defensive act. Every learned detector is a differentiable or approximable function that an adversary may optimise against, and the corpus contains no instance of a mechanism evaluated under that assumption. The mechanisms most exposed are those whose representations are lowest in abstraction, since these are precisely the representations an attacker can perturb most cheaply – which returns the analysis to Finding 1 and suggests that abstraction level and adversarial robustness are two descriptions of a single underlying property.

4.2.4. Anonymity, accountability and their reconciliation

S4 addresses a tension that the assurance literature more often manages than resolves: strong anonymity for receivers appears to conflict with accountability for senders, since a system in which no party can be identified is a system in which no party can be held responsible. The scheme's resolution is architectural. Receiver identities are removed from the ciphertext entirely and replaced by randomised index values from which identity cannot be recovered, while sender accountability is preserved through un-

Table 15. Summary of findings and certainty of the evidence.

#	Finding	Contributing studies	Nature of supporting evidence	Certainty	Principal reason for the rating
F1	Assurance mechanisms converge on an invariance principle, and the abstraction level of the feature representation is the dominant determinant of robustness.	S1, S2, S3, S4	Explicit taxonomy in one study; design rationales consistent with the principle in three others	Moderate	Consistent and direct, but each layer is represented by a single study, so the pattern rests on breadth rather than replication
F2	Provable and empirical assurance are asymmetrically distributed: one study in five supplies formal reduction, four rely on benchmark performance alone.	S1, S2, S3, S4, S5	Direct property of the extracted design characteristics; requires no inference	High	Directly observed in every included study with no inconsistency and no imprecision
F3	Evaluation infrastructure is mature at two layers, emerging at one and absent at two.	S1, S2, S3, S4, S5	Extracted dataset characteristics, corroborated by explicit statements from the primary authors in two studies	Moderate to high	Direct and author-corroborated; downgraded slightly because maturity is judged rather than measured
F4	Adaptive adversaries are documented within the corpus but evaluated by none of the included studies.	S1 documents the threat; S2, S3 and S5 do not evaluate it	Documented threat in one study; verifiable absence from the evaluation design of the others	Moderate	The absence is directly verifiable, but the magnitude of the resulting overstatement of assurance cannot be estimated from this corpus
F5	Where quantified, the cost of assurance is modest relative to the assurance obtained, but cost quantification is itself inconsistent.	S2, S3, S4, S5	Analytically derived overhead in one study, asymptotic complexity in another, measured trade-offs in two	Low to moderate	Individual measurements sound, but cost dimensions are not commensurable and two studies report no dispersion

forgeability together with a controlled disclosure path by which a key-generation authority may, on a legitimate request, determine the sender's identity.

Read against S1, this resolution acquires additional interest. Among the alternative approaches that survey documents are watermarking of control-flow structures, embedded developer identifiers bound to signing keys, and authorship attribution from coding style – all mechanisms whose purpose is precisely to establish accountability for a software artefact, and one of which is reported as attributing authorship among hundreds of developers at accuracy exceeding eighty per cent. The two studies are therefore addressing opposite faces of the same problem: S4 is concerned to prevent identity inference, S1 documents work concerned to enable it. That both appear in a single volume of a single journal, without cross-reference, is a compact illustration of the fragmentation this review was designed to expose.

4.3. CROSS-CUTTING THEMES

4.3.1. The cost of assurance is architectural, not merely computational

Every study that quantifies cost locates it structurally rather than treating it as an undifferentiated overhead. S3 distinguishes three

phases and observes that only the third generates network traffic, since propagation estimation is centralised and periodic while local detection runs on-device; the consequence is that the overhead scales with neighbour density but is independent of network size for a fixed density-to-hop-distance ratio, which is an architectural rather than an algorithmic property. S4 separates transmission cost from computation cost and finds its advantage in the growth rate of the former. S2 separates encoder from decoder complexity and observes that the asymmetry is deliberate, the larger decoder purchasing extraction robustness while the lighter encoder preserves embedding efficiency. S5 separates search cost from inference cost, the former being incurred once and the latter repeatedly.

The practical implication is that assurance costs should be reported as a profile rather than a scalar, since the profile determines deployability. A mechanism with high one-off cost and negligible recurring cost is deployable on constrained hardware; a mechanism with the same total cost distributed recurrently is not. None of the studies states this principle in general terms, but all four that measure cost observe it in practice.

Table 16. Research agenda: gaps, directions and traceability.

#	Direction	Gap addressed	Source	Tier
R1	Construct an open benchmark corpus for physical-layer authentication in metering and sensor networks, containing bidirectional signal-strength measurements between all node pairs with synchronised environmental measurements and both genuine and adversarially generated impersonation events across urban, suburban and rural morphologies.	No corpus exists with the required conjunction of properties; evaluation consequently rests on augmented data with injected attacks and cannot be corroborated independently.	S3 (conceded); §3.6.1	1
R2	Establish a standard cost model and reference implementation environment for comparative evaluation of signcryption and related public-key primitives, specifying operation-level accounting and a common curve, library and platform, so that efficiency claims do not depend on each author's reimplementation of competitors.	No benchmark or standard cost model exists at the protocol layer; comparison rests on self-reimplementation of comparator schemes.	S4; §3.6.1	1
R3	Evaluate learning-based assurance mechanisms against adaptive adversaries who optimise transformations specifically against the deployed model, adopting the problem-space attack methodology as a standard evaluation condition rather than as a separate research topic.	No included study evaluates under adaptive attack, although the corpus itself documents such attacks succeeding at scale against both academic and commercial detectors.	S1 (documents the threat); §3.4, §4.2.3	1
R4	Develop and validate a methodology for comparing assurance guarantees of different epistemic kinds, permitting a conditional asymptotic guarantee and an empirical detection rate to be placed on a common footing for the purpose of reasoning about composite systems.	Guarantees of incommensurable kinds are routinely composed in deployed systems with no principled basis for identifying the binding constraint.	S3 with S4; §4.2.2	1
R5	Adopt dispersion reporting as standard practice for stochastic assurance mechanisms, requiring repeated runs with standard deviations and confidence intervals wherever training or search involves randomness.	Only one of four empirical studies reports variability; reported margins over baselines cannot be evaluated against run-to-run variation.	S2, S3, S4; §3.3.3	2
R6	Extend third-party library identification from library level to version level with robustness to modern code-shrinking and optimisation pipelines, and construct version-annotated ground truth at a scale sufficient to support it.	Version-level discrimination is identified as a critical bottleneck notwithstanding high library-level accuracy.	S1; §3.8	2
R7	Integrate watermark embedding with content synthesis rather than applying it post hoc, and add spatial or temporal localisation so that the region of a tampered object can be identified rather than only the fact of tampering detected.	The included watermarking scheme is post hoc and provides no localisation, which its authors concede.	S2 (conceded); §3.8	2
R8	Construct signcryption schemes providing sender anonymity jointly with receiver anonymity while preserving non-repudiation through a controlled disclosure mechanism, and prove them secure in the standard rather than the random-oracle model.	Only receiver anonymity is achieved; sender anonymity is conceded as absent and proofs are confined to the random-oracle model.	S4 (conceded); §3.8	2

4.3.2. Cooperation and redundancy as assurance amplifiers

Two studies exploit redundancy across independent observers, and both report substantial gains. S3's cooperative phase raises

precision from 0.7500 to 0.9380 while preserving recall, on the reasoning that a malicious node may deceive one observer but is unlikely to deceive several simultaneously from a single physical

Table 16. Research agenda: gaps, directions and traceability(continued).

#	Direction	Gap addressed	Source	Tier
R9	Formulate architecture search for leakage analysis as an explicitly multi-objective problem over attack effectiveness, model complexity and search cost, rather than optimising effectiveness alone and reporting complexity descriptively.	Parameter count is measured but not optimised; the authors identify this as an extension.	S5 (conceded); §3.7	2
R10	Investigate cooperative and redundancy-based verification as a general assurance amplifier at layers where it is not currently used, in particular for software-artefact and media-object verification, and characterise the conditions under which independence of observers can be assumed.	Cooperative verification produces substantial gains at one layer and appears in hybrid form at another, but is not treated as a general design pattern.	S3, S1; §4.3.2	3
R11	Establish co-evolutionary evaluation protocols in which detection and evasion are improved alternately over multiple rounds, and report assurance as a function of adversary adaptation budget rather than as a single-round figure.	Assurance is uniformly reported against static adversaries, giving no information about durability under sustained adaptation.	S1, S5; §4.3.4	3

position. S1 documents an analogous logic in the multi-feature and hybrid detection approaches it surveys, where combining code, interface, resource and library evidence is repeatedly reported to improve robustness relative to any single view, at the price of computational complexity.

The underlying principle is that independence of evidence sources converts a single-point decision into an agreement problem, and that forging agreement across independent sources is disproportionately harder than forging any one of them. This is the same principle that underlies threshold cryptography and multi-sensor fusion, and its independent appearance at two unrelated layers of this corpus suggests it is among the more transferable ideas available. Its limits are equally visible: S3's degraded configuration, in which only one neighbour is available, loses most of the gain, which establishes that the amplification depends on genuine observational redundancy and cannot be manufactured where redundancy is absent.

4.3.3. Benchmark availability determines the shape of a field

The corpus supplies an unusually clean natural comparison. Where mature public benchmarks exist, the studies using them are able to report comparisons against retrained baselines, to test cross-domain generalisation on held-out corpora, and to compute dispersion statistics over many runs; S2 and S5 both do so, and both achieve the highest quality scores among the empirical studies. Where benchmarks are absent, the corresponding practices are absent too: S3 must construct its evaluation and reports no dispersion, and S4 must reimplement its competitors to compare against them.

The direction of causation cannot be established from five studies, and it is entirely plausible that mature fields produce both benchmarks and rigorous practice for a common underlying reason. What can be said is that the association is present in every case in this corpus, and that the two fields lacking benchmarks are also the two in which the primary authors themselves identify evaluation as the binding constraint. This is a strong *prima facie*

argument for benchmark construction as the highest-leverage investment in those areas, and it is the basis of the first two items of the research agenda in Section 6.

4.3.4. The evaluative literature bounds the defensive literature

S5 occupies a distinctive position, being the only study whose success constitutes a reduction rather than an increase in assurance. Its inclusion changes the reading of the corpus as a whole. Read defensively, the corpus reports a series of mechanisms performing well against the adversaries they model. Read with S5 included, it reports in addition that automated search can be used to make attacks against cryptographic implementations markedly more efficient – recovering a masked key byte within approximately sixty traces where conventional search strategies require several times as many – and that the same class of technique is reported in the wider literature as capable of recovering architectural information about a neural accelerator from its physical leakage [36].

The methodological lesson generalises. Any assessment of an assurance ecosystem that samples only defensive contributions will systematically overstate the assurance available, because the evaluative literature establishes the ceiling and the defensive literature reports progress toward it. The eligibility criteria of this review were framed to admit evaluative work for precisely this reason, and the resulting reading is materially different from what a defensively scoped review would have produced.

4.4. SENSITIVITY OF THE SYNTHESIS

Each cross-cutting theme was re-derived with each study omitted in turn, in order to establish whether any theme depended on a single study. The results are as follows. The invariance principle of Section 4.3 survives the omission of any single study, being independently evidenced in S1, S2, S3 and S4, although it is stated most explicitly in S1 and loses its clearest articulation when that study is removed. The theme of architectural cost survives the omission of any single study, being evidenced in four. The co-

operation theme depends substantially on S3 and is weakened to a suggestion drawn from S1's survey material alone when S3 is omitted. The benchmark theme is weakened but survives, since the primary authors' own statements in S1 and S3 each independently support it. The evaluative-ceiling theme depends entirely on S5 and does not survive its omission.

This last dependency is material and is reported as such. Finding 4 and the argument of Section 4.3.4 rest on a single study together with material that S1 reports at second hand. They should accordingly be read as the most provisional conclusions of this review, notwithstanding that S5 received the highest quality score in the corpus.

5. LIMITATIONS OF THIS REVIEW

5.1. LIMITATIONS ARISING FROM THE SAMPLING FRAME

The most consequential limitation of this review is the bounded sampling frame. Restricting identification to a single volume of a single journal means that the review does not, and does not claim to, constitute an exhaustive census of the world literature on assurance mechanisms. Four specific consequences follow and are stated without mitigation.

First, generalisability is limited. The five included studies are not a random or representative sample of contemporary work; they are the population of one venue in one period. Mechanisms, layers and methodological practices well represented elsewhere may be absent here, and their absence from this corpus carries no implication about their prevalence in the field. Where the review reports that a practice is rare, that statement is a statement about this corpus and must not be read as a statement about the literature.

Second, venue effects cannot be separated from field effects. The editorial preferences, review standards and topical emphases of a single journal are confounded with the properties of the research areas sampled. The observation that provable-security contributions are outnumbered four to one by empirical ones, for instance, may reflect the journal's applied orientation rather than any property of the field.

Third, statistical power is negligible. With five studies and no within-layer replication, no quantitative claim about the distribution of practices can be supported, and none is made. All numerical statements in Section 3 are descriptions of individual studies, not estimates of population parameters.

Fourth, the term 'systematic' requires qualification. The review is systematic in the sense that matters most for validity – the criteria, procedures, extraction fields and appraisal rubric were pre-specified, applied uniformly and reported in full, so that any reader may verify or contest every judgement. It is not systematic in the sense of comprehensive retrieval. Readers requiring exhaustive coverage of any one of the five layers should consult a topically scoped review of that layer, and S1 is itself an example of what such a review looks like for the application layer.

The countervailing consideration, stated once and not overstated, is that this design purchases a property that broad reviews rarely achieve: complete verifiability of the corpus and item-level appraisal of every included study. Whether that trade is worthwhile depends on the use to which the review is put; it is well suited to establishing structural relationships across layers and poorly suited to establishing prevalence.

5.2. LIMITATIONS OF PROCESS

Selection, extraction and appraisal were performed independently and in duplicate with a pre-specified adjudication procedure, which is the standard safeguard against reviewer error. Duplicate assessment nevertheless provides no protection against a misconception shared by both reviewers, and the reviewers in this case worked within a single research group and were therefore not independent in their disciplinary formation. Complete agreement between reviewers, as obtained here, is consistent with correct and unambiguous application of the criteria, but it is equally consistent with a shared misreading, and the two possibilities cannot be distinguished from within the review. Item-level publication of all decisions and scores is offered as the mitigation, since it transfers the possibility of correction to the reader.

The appraisal rubric was constructed for this review rather than adopted from a validated instrument. It has not been assessed for inter-rater reliability, and its domain weights are implicitly equal, which is a substantive assumption rather than an established fact. A reader who considers, for example, that analytical rigour should be weighted more heavily than limitation reporting would obtain different aggregate scores from the same item-level data, which is why the item-level data are published.

Certainty assessment was narrative rather than instrumented. No validated framework for rating certainty in bodies of computational-security evidence exists, and the four considerations adapted from clinical practice are applied here by judgement rather than by algorithm.

5.3. DEVIATIONS FROM THE PROTOCOL

One deviation occurred. The protocol specified that the quality appraisal would score every domain for every study on the three-point scale. During appraisal it became apparent that the dataset domain is genuinely inapplicable to a purely theoretical construction, and that scoring it zero would penalise genre rather than quality. The rubric was therefore amended to permit a 'not applicable' code excluded from both numerator and denominator, and the amendment was applied uniformly and retrospectively to all studies. The amendment affected one cell, in S4, and raised that study's reported score from 81.3 per cent to 92.9 per cent. It is reported here so that readers may substitute their own judgement if they consider the original specification preferable. A supplementary clarification is warranted, since Study S4 reports execution timings on both a desktop workstation and an embedded platform. These hardware measurements are appraised under Q5 (Methods and Analysis) and Q6 (Comparator Adequacy) rather than under Q4 (Data/Benchmark). Q5 governs the transparency and appropriateness of the timing procedure itself – the hardware configuration, the number of repetitions and the operation-cost notation – while Q6 governs the appropriateness of the baseline against which those timings are set, in this case the reimplementation of three published comparator schemes on identical hardware. Q4, by construction of the rubric, is reserved for the use of an external dataset or benchmark corpus that a study neither owns nor generates; a signcryption construction whose evaluation object is the scheme itself has no such corpus, and none is required by its design. This is the ground on which Q4 was coded 'not applicable' for S4, and it is the substantive difference between measuring the implementation and measuring against a shared



Figure 5. Evidence gap map across the five assurance layers and seven dimensions of evidential adequacy. Ratings are derived from the extracted data of Tables 8–14 and follow the appraisal recorded in Appendix C.

corpus.

5.4. LIMITATIONS OF THE EVIDENCE BASE ITSELF

Distinct from the limitations of the review are limitations in what the included studies collectively can support. Three are material. There is no within-layer replication, so no finding is corroborated by an independent study of the same mechanism at the same layer. Dispersion statistics are reported by only one study, so most reported differences cannot be evaluated against run-to-run variability. And no study evaluates against an adaptive adversary optimising against the deployed mechanism, so the corpus provides no evidence at all about performance under the threat model that Section 4.2.3 identifies as most consequential.

6. RESEARCH AGENDA

The agenda below is derived by mapping each evidential gap identified in Section 3 to a corresponding research direction. Every item is traceable to a specific gap and to the studies that evidence it; no item is included on general grounds. Items are grouped into three tiers by the review’s assessment of leverage, where leverage is judged by the number of downstream findings a resolution would unblock.

Figure 5 presents the same analysis as an evidence gap map, in which each assurance layer is assessed against seven dimensions of evidential adequacy. The map makes visible two patterns that the tabular agenda does not. First, the deficiency in within-layer replication is universal: no layer in this corpus is served by more than one study, so the rightmost column is empty throughout. Second, the gaps are not randomly distributed but cluster by layer, with the network and protocol layers deficient chiefly in benchmark infrastructure and the media and application layers deficient chiefly in adversarial evaluation. Interventions targeted by layer are therefore likely to be more efficient than interventions targeted by dimension.

Two observations about the agenda as a whole. First, three of

the four tier-one items concern evaluation infrastructure rather than mechanism design, which reflects the review’s assessment that the binding constraint in this domain is currently the capacity to establish what mechanisms achieve rather than the capacity to devise them. Second, the fourth tier-one item, concerning the comparison of guarantees of different kinds, is the only item that no individual study could have identified, since it arises solely from the juxtaposition of studies at different layers. It is therefore the item most specific to the contribution of a cross-layer synthesis.

7. CONCLUSION

This review set out to determine how contemporary research assures the authenticity, integrity and provenance of digital artefacts across the layers of the computing stack, and to appraise the evidence supporting the assurance claimed. Five studies drawn from a defined and exhaustively enumerated sampling frame were screened, appraised on an eight-domain rubric and synthesised narratively against five review questions. The included studies occupy five distinct layers – the application artefact, the media object, the networked device, the cryptographic protocol and the micro-architectural implementation – and their methodological quality ranged from 81.3 to 93.8 per cent of the attainable score.

The synthesis yields three conclusions of substance. The first is that assurance mechanisms, however different their objects and mathematics, converge on a common principle: they seek representations invariant under expected transformation and sensitive to adversarial modification, and the level of abstraction at which such a representation is formed is the strongest single determinant of its durability. Token-level representations fall to reordering; structural representations survive it; semantic and algebraic representations survive more still. This principle organises the technical content of the corpus and predicts the direction in which the arms race at each layer has in fact moved.

The second is that the assurance claimed is not everywhere of the same kind. One study in five discharges its claims by reduction to stated hardness assumptions; the remainder discharge theirs by benchmark performance. Both are legitimate, but they are not commensurable, and deployed systems routinely compose them without any principled means of identifying which guarantee binds. Developing such a means is, on the evidence of this review, the most distinctively cross-layer problem the field faces.

The third is that the adversary the corpus describes is not the adversary against which it tests. Adaptive, evasion-optimising attackers are documented within the corpus itself as succeeding at scale against both academic and commercial detectors, yet no study evaluates its own mechanism under that condition. Until adaptive evaluation becomes routine rather than exceptional, reported assurance figures should be read as upper bounds obtained against static adversaries, and the margin between those bounds and the assurance available in adversarial deployment remains unquantified.

These conclusions are advanced with the caution that the review's design requires. The sampling frame is bounded, the corpus contains no within-layer replication, and the review – including screening, extraction and quality appraisal – was executed independently and in duplicate by two reviewers, with a pre-specified third-reviewer adjudication procedure that in the event was not required, as documented in Sections 2.5, 2.6, 2.7 and 5.2. The findings are accordingly offered as a structural account of how a defined body of contemporary evidence is composed and what it can and cannot support, and as a prioritised agenda of eleven directions each traceable to a specific deficiency in that evidence. Three of the four highest-priority directions concern the infrastructure of evaluation rather than the design of mechanisms – which is itself the review's most compact finding, and the one on which progress at every layer appears to depend.

DATA AVAILABILITY

All data analysed in this review were extracted from the published full texts of the five included studies, each of which is openly accessible under a Creative Commons licence at the digital object identifiers listed in the reference list. The completed extraction form, the screening decision record and the item-level appraisal rationale are reproduced in Appendices A to C, so that the analysis is reconstructible from this document together with the primary sources. No new empirical data were generated.

DECLARATION OF COMPETING INTEREST

The authors declare no financial or non-financial competing interests in relation to the subject matter of this review. The authors have no affiliation with, and no prior collaboration with, the authors of any included study, and no involvement in the editorial processes of the journal constituting the sampling frame.

FUNDING

This review received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

ACKNOWLEDGMENT

The authors thank the Department of Mathematics and Computer Science, Arthur Jarvis University, Akpabuyo, and the Mathematics Programme, Department of Physical Sciences, Landmark University, Omu-Aran, for institutional support. The authors further acknowledge the authors of the five included studies, whose decision to publish under open-access licences made the item-level appraisal reported here possible.

References

- [1] Y. Wu, Z. Liu, Y. Lin, B. Zhao, C. Zhou, T. Wu & Z. Hong, "Android app repackaging detection: a comprehensive survey", *Cyber Security and Applications* 4 (2026) 100124. <https://doi.org/10.1016/j.csa.2026.100124>.
- [2] S. Wen, H. Tan, J. Li & T. Hu, "A robust audio watermarking method based on dual-encoder U-Net and short-time Fourier transform", *Cyber Security and Applications* 4 (2026) 100129. <https://doi.org/10.1016/j.csa.2026.100129>.
- [3] B. Achaal, M. Adda, M. Berger & A. Awde, "An intelligent authentication and intrusion detection system for securing advanced metering infrastructure", *Cyber Security and Applications* 4 (2026) 100115. <https://doi.org/10.1016/j.csa.2025.100115>.
- [4] Y. Zhao, X. Liu, Y. Zhou, B. Yang & M. Zhang, "A multi-receiver and multi-message anonymous ID-based hybrid signcryption scheme for secure data transmission", *Cyber Security and Applications* 4 (2026) 100128. <https://doi.org/10.1016/j.csa.2026.100128>.
- [5] A. Amrouche, L. Boubchir & S. Yahiaoui, "Metaheuristic-driven neural architecture search for deep learning-based side-channel analysis", *Cyber Security and Applications* 4 (2026) 100125. <https://doi.org/10.1016/j.csa.2026.100125>.
- [6] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting & D. Moher, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews", *BMJ* 372 (2021) n71. <https://doi.org/10.1136/bmj.n71>.
- [7] H. Li, Z. Yao, B. Wu, C. Gao, T. Xu, W. Yuan & X. Luo, "Automated mass malware factory: the convergence of piggybacking and adversarial example in Android malicious software generation", *Proceedings of the 32nd Network and Distributed System Security Symposium (NDSS)*, San Diego, USA, 2025. <https://doi.org/10.14722/ndss.2025.241933>.
- [8] L. Li, T. F. Bissyandé & J. Klein, "Rebooting research on detecting repackaged Android apps: literature review and benchmark", *IEEE Transactions on Software Engineering* 47 (2021) 676. <https://doi.org/10.1109/TSE.2019.2901679>.
- [9] X. Zhan, T. Zhang & Y. Tang, "A comparative study of Android repackaged apps detection techniques", *2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, Hangzhou, China, 2019, pp. 321–331. <https://doi.org/10.1109/SANER.2019.8667975>.
- [10] W. Zhou, Y. Zhou, X. Jiang & P. Ning, "Detecting repackaged smartphone applications in third-party Android marketplaces", *Proceedings of the Second ACM Conference on Data and Application Security and Privacy (CO-DASPY)*, San Antonio, USA, 2012, pp. 317–326. <https://doi.org/10.1145/2133601.2133640>.
- [11] K. Allix, T. F. Bissyandé, J. Klein & Y. Le Traon, "AndroZoo: collecting millions of Android apps for the research community", *Proceedings of the 13th International Conference on Mining Software Repositories (MSR '16)*, Austin, USA, 2016, pp. 468–471. <https://doi.org/10.1145/2901739.2903508>.
- [12] T. S. John, T. Thomas & S. Emmanuel, "Detection of evasive Android malware using EigenGCN", *Journal of Information Security and Applications* 86 (2024) 103880. <https://doi.org/10.1016/j.jisa.2024.103880>.
- [13] L. Zhan, J. Ming, J. Fu, G. Peng, L. Sha & L. Lan, "The hidden complexities of Android TPL detection: an empirical analysis of techniques, challenges, and effectiveness", *Computers & Security* 159 (2025) 104672. <https://doi.org/10.1016/j.cose.2025.104672>.
- [14] Z. Zhang, S. Luo, Y. Lu & L. Pan, "Obfuscation-resilient detection of Android third-party libraries using multi-scale code dependency fusion", *Information Fusion* 117 (2025) 102908. <https://doi.org/10.1016/j.inffus.2024.102908>.
- [15] Q. Zeng, L. Luo, Z. Qian, X. Du & Z. Li, "Resilient decentralized An-

- droid application repackaging detection using logic bombs”, Proceedings of the 2018 International Symposium on Code Generation and Optimization (CGO), Vienna, Austria, 2018, pp. 50–61. <https://doi.org/10.1145/3168820>.
- [16] G. Lin, W. Luo, P. Zheng & J. Huang, “An audio watermarking method against re-recording distortions”, *Pattern Recognition* **162** (2025) 111366. <https://doi.org/10.1016/j.patcog.2025.111366>.
- [17] C. Liu, J. Zhang, H. Fang, Z. Ma, W. Zhang & N. Yu, “DeAR: a deep-learning-based audio re-recording resilient watermarking”, Proceedings of the AAAI Conference on Artificial Intelligence **37** (2023) 13201. <https://doi.org/10.1609/aaai.v37i11.26550>.
- [18] C. Liu, J. Zhang, T. Zhang, X. Yang, W. Zhang & N. Yu, “Detecting voice cloning attacks via timbre watermarking”, arXiv preprint arXiv:2312.03410 (2023). Available online: <https://arxiv.org/abs/2312.03410>.
- [19] G. Hua, J. Huang, Y. Q. Shi, J. Goh & V. L. L. Thing, “Twenty years of digital audio watermarking: a comprehensive review”, *Signal Processing* **128** (2016) 222. <https://doi.org/10.1016/j.sigpro.2016.04.005>.
- [20] F. T. Liu, K. M. Ting & Z.-H. Zhou, “Isolation Forest”, Proceedings of the IEEE International Conference on Data Mining (ICDM), Pisa, Italy, 2008, pp. 413–422. <https://doi.org/10.1109/ICDM.2008.17>.
- [21] A. Pourghasem, R. Kirner, A. Tsokanos, I. Mporas & A. Mylonas, “Machine learning and deep learning-based multi-attribute physical-layer authentication for spoofing detection in LoRaWAN”, *Future Internet* **17** (2025) 68. <https://doi.org/10.3390/fi17020068>.
- [22] Z. Yun & M. F. Iskander, “Ray tracing for radio propagation modeling: principles and applications”, *IEEE Access* **3** (2015) 1089. <https://doi.org/10.1109/ACCESS.2015.2453991>.
- [23] B. Achaal, M. Adda, M. Berger, H. Ibrahim & A. Awde, “Study of smart grid cybersecurity: examining architectures, communication networks, cyber-attacks, countermeasure techniques and challenges”, *Cybersecurity* **7** (2024) 10. <https://doi.org/10.1186/s42400-023-00200-w>.
- [24] Y. Zheng, “Digital signcryption or how to achieve $\text{cost}(\text{signature and encryption}) \leq \text{cost}(\text{signature}) + \text{cost}(\text{encryption})$ ”, in *Advances in Cryptology – CRYPTO ’97*, B. S. Kaliski (Ed.), Springer, Berlin, Germany, 1997, pp. 165–179. <https://doi.org/10.1007/BFb0052234>.
- [25] R. Cramer & V. Shoup, “Design and analysis of practical public-key encryption schemes secure against adaptive chosen ciphertext attack”, *SIAM Journal on Computing* **33** (2003) 167. <https://doi.org/10.1137/S0097539702403773>.
- [26] A. W. Dent, “Hybrid signcryption schemes with outsider security”, in *Information Security – ISC 2005*, J. Zhou, J. Lopez, R. H. Deng & F. Bao (Eds.), Springer, Berlin, Germany, 2005, pp. 203–217. https://doi.org/10.1007/11556992_15.
- [27] Y. Zhou, R. Xu, Z. Qiao, B. Yang, Z. Xia & M. Zhang, “An anonymous and efficient multimesage and multireceiver certificateless signcryption scheme for VANET”, *IEEE Internet of Things Journal* **10** (2023) 22823. <https://doi.org/10.1109/JIOT.2023.3305489>.
- [28] Y. Liang, H. Yan & Y. Liu, “Unlinkable signcryption scheme for multi-receiver in VANETs”, *IEEE Transactions on Intelligent Transportation Systems* **24** (2023) 10138. <https://doi.org/10.1109/TITS.2023.3271110>.
- [29] L. Wu, G. Perin & S. Picek, “I choose you: automated hyperparameter tuning for deep learning-based side-channel analysis”, *IEEE Transactions on Emerging Topics in Computing* **12** (2024) 546. <https://doi.org/10.1109/TETC.2022.3218372>.
- [30] S. Picek, G. Perin, L. Mariot, L. Wu & L. Batina, “SoK: deep learning-based physical side-channel analysis”, *ACM Computing Surveys* **55** (2023) 227:1. <https://doi.org/10.1145/3569577>.
- [31] H. Maghrebi, T. Portigliatti & E. Prouff, “Breaking cryptographic implementations using deep learning techniques”, in *Security, Privacy and Applied Cryptography Engineering (SPACE 2016)*, C. Carlet, M. A. Hasan & V. Saraswat (Eds.), Springer, Cham, Switzerland, 2016, pp. 3–26. https://doi.org/10.1007/978-3-319-49445-6_1.
- [32] P. Kocher, J. Jaffe & B. Jun, “Differential power analysis”, in *Advances in Cryptology – CRYPTO ’99*, M. Wiener (Ed.), Springer, Berlin, Germany, 1999, pp. 388–397. https://doi.org/10.1007/3-540-48405-1_25.
- [33] S. Chari, J. R. Rao & P. Rohatgi, “Template attacks”, in *Cryptographic Hardware and Embedded Systems – CHES 2002*, B. S. Kaliski, Ç. K. Koç & C. Paar (Eds.), Springer, Berlin, Germany, 2003, pp. 13–28. https://doi.org/10.1007/3-540-36400-5_3.
- [34] T. Elsken, J. H. Metzen & F. Hutter, “Neural architecture search: a survey”, *Journal of Machine Learning Research* **20** (2019) 1. Available online: <https://www.jmlr.org/papers/v20/18-598.html>.
- [35] A. Gohr, S. Jacob & W. Schindler, “CHES 2018 side channel contest CTF: solution of the AES challenges”, *Cryptology ePrint Archive*, Report 2019/094, 2019. Available online: <https://eprint.iacr.org/2019/094>.
- [36] N. Gupta, A. Jati & A. Chattopadhyay, “AI attacks AI: recovering neural network architecture from NVDLA using AI-assisted side-channel attack”, *ACM Transactions on Embedded Computing Systems* **24** (2025) 75. <https://doi.org/10.1145/3731560>.
- [37] M. Defferrard, K. Benzi, P. Vandergheynst & X. Bresson, “FMA: a dataset for music analysis”, arXiv preprint arXiv:1612.01840 (2016). Available online: <https://arxiv.org/abs/1612.01840>.
- [38] V. Panayotov, G. Chen, D. Povey & S. Khudanpur, “LibriSpeech: an ASR corpus based on public-domain audio books”, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 2015, pp. 5206–5210. <https://doi.org/10.1109/ICASSP.2015.7178964>.
- [39] P. Tian, F. Yang, X. Ma, C. A. Boano, X. Tian, Y. Liu & J. Wei, “Dataset: environmental impact on the long-term connectivity and link quality of an outdoor LoRa network”, Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems (SenSys), Coimbra, Portugal, 2021, pp. 565–568. <https://doi.org/10.1145/3485730.3493696>.
- [40] F. Pierazzi, F. Pendlebury, J. Cortellazzi & L. Cavallaro, “Intriguing properties of adversarial machine-learning attacks in the problem space”, 2020 IEEE Symposium on Security and Privacy (S&P), San Francisco, USA, 2020, pp. 1332–1349. <https://doi.org/10.1109/SP40000.2020.00073>.

APPENDIX A. SEARCH STRATEGY AND ENUMERATION RECORD

The sampling frame was enumerated rather than queried (see Section 2.4), and the enumeration cutoff of 15 September 2026 is the date against which the Boolean expression below is calibrated. For readers wishing to reconstruct an equivalent corpus from a general bibliographic index, the Boolean expression below, applied with the stated limits and constrained to records available on or before the cutoff date, would retrieve a set materially equivalent to the enumerated frame. It is recorded to satisfy the replicability requirement of the reporting guideline, not because it was the retrieval method actually used, and readers seeking to reproduce the exact frame should use enumeration rather than query, since a database search cannot guarantee identity of coverage with the journal volume itself.

APPENDIX B. SCREENING AND ELIGIBILITY DECISION RECORD

Every screening and eligibility decision is recorded below with the criterion invoked, in accordance with Section 2.5. Reviewer 1 and Reviewer 2 denote the two reviewers who screened independently and blind to one another; the adjudication column records the determination of the third reviewer where referral was required. Agreement between the two independent reviewers was complete at both the title/abstract stage (nine records, four excluded and five retained) and the full-text stage (five records, all retained). No referral to the third reviewer was required. The four exclusions at the title/abstract stage were on Criterion I of Table 1: the article did not propose a detective, constructive or evaluative mechanism for authenticity, integrity or provenance.

Chance-corrected agreement is reported as follows. At the title/abstract stage the marginal distribution comprises two categories (excluded and retained), and Cohen’s kappa is computable on that stage; observed agreement was 100 per cent (nine of nine records classified identically), so kappa is 1.00, corresponding to perfect agreement. At the full-text stage every record was classified identically as retained and no record fell into any other

category; Cohen's kappa is therefore undefined for the full-text stage, since the coefficient requires variation in at least two categories to be computable, and raw percentage agreement of 100 per cent is reported instead. Complete agreement on five full-text records carries limited information about reliability in either direction, and no claim of demonstrated inter-rater reliability is advanced on this basis; item-level publication of all decisions in Appendix B is offered as the mitigation.

In accordance with the convention that a systematic review should report the characteristics of excluded studies, the four records excluded at title/abstract screening are listed in Table B1 below with the criterion invoked and the article identifier. No record was excluded on grounds of language, availability, publication status or date.

APPENDIX C. QUALITY APPRAISAL: SCORING RATIONALE

The rationale for every non-maximal score is given below, together with the rationale for the single 'not applicable' code. Scores of 2 not addressed here were awarded because the domain's anchor descriptor in Table 4 was met without qualification. Aggregate scores were computed as the sum of applicable domain scores divided by twice the number of applicable domains, expressed as a percentage. For S4 the denominator is 14 rather than 16. The Study S4 execution timings on desktop and embedded hardware are scored under Q5 (Methods and Analysis) and Q6 (Comparator Adequacy) – because they concern the transparency of the timing procedure and its comparison against three reimplemented baseline schemes – rather than under Q4, which is reserved for the use of an external dataset or benchmark corpus. No such corpus is applicable to a signcryption construction whose evaluation object is the scheme itself, and the Q4 code of 'not applicable' for S4 is on that ground.

APPENDIX D. ABBREVIATIONS

AES Advanced Encryption Standard

AMI Advanced metering infrastructure

ASCAD ANSSI SCA Database (side-channel benchmark corpus)

AUC Area under the receiver-operating characteristic curve

CDH Computational Diffie–Hellman (problem or assumption)

CHES Conference on Cryptographic Hardware and Embedded Systems

CNN Convolutional neural network

DBDH Decisional bilinear Diffie–Hellman (problem or assumption)

DEM Data encapsulation mechanism

ERSSI Expected received signal strength indicator

EUFCMA Existential unforgeability under adaptive chosen-message attack

GE Guessing entropy

GRADE Grading of Recommendations, Assessment, Development and Evaluations

HW Hamming weight (leakage model)

ID Identity (leakage model); or identity-based (cryptography)

IND-CCA Indistinguishability under chosen-ciphertext attack

KEM Key encapsulation mechanism

MAE Mean absolute error

MLP Multilayer perceptron

MRA-IBHS Multi-receiver anonymous identity-based hybrid signcryption

NAS Neural architecture search

PICOS Population, Intervention, Comparator, Outcomes, Study design

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RMSE Root mean square error

RSSI Received signal strength indicator

SCA Side-channel analysis

STFT Short-time Fourier transform

TPL Third-party library

APPENDIX E. PRISMA 2020 CHECKLIST

The location of every item of the PRISMA 2020 reporting guideline is recorded below. Where an item is reported as not applicable, the reason is stated, since silent omission would be indistinguishable from oversight.

BOOLEAN EXPRESSION (SCOPUS SYNTAX)

(ISSN(2772-9184) OR SRCTITLE("Cyber Security and Applications"))

AND VOLUME(4)

AND PUBYEAR = 2026

AND DOCTYPE(ar OR re)

AND LANGUAGE(english)

(Equivalent Web of Science syntax: IS=2772-9184 AND PY=2026 AND VL=4 AND DT=(Article OR Review) AND LA=English. Equivalent Dimensions syntax: source.issn="2772-9184" and year=2026 and volume=4 and type in ("article", "review")).

Note of record: As stated in Section 2.4 and Appendix A, this specification is recorded to satisfy the replicability requirement of the reporting guideline; it was not the retrieval method actually used. Readers seeking to reproduce the exact frame should enumerate the volume rather than query, since no bibliographic index can guarantee identity of coverage with the journal volume itself at an arbitrary cutoff.

Table A1. Equivalent retrieval specification.

Parameter	Value
Target bibliographic index	Scopus (equivalent specification available for Web of Science Core Collection and Dimensions)
Source identifier	ISSN 2772-9184 (print and electronic); publisher: Elsevier B.V. on behalf of KeAi Communications
Source title	Cyber Security and Applications
Field codes used	SRCTITLE (source title); ISSN; PUBYEAR (year of publication); VOLUME; DOCTYPE (document type); LANGUAGE
Document-type limit	Peer-reviewed research articles and peer-reviewed reviews (DOCTYPE: ar OR re)
Language limit	English (LANGUAGE: english)
Volume limit	VOLUME = 4
Year limit	PUBYEAR = 2026
Date-of-availability limit	Records with online availability on or before the enumeration cutoff of 15 September 2026
DOI prefix constraint (optional)	DOI matches 10.1016/j.csa.*
Supplementary methods	None. Citation chasing (forward or backward) and grey-literature retrieval were excluded by design (see Section 2.4).
Expected yield at cutoff	Nine records (article numbers 100115, 100124, 100125, 100126, 100127, 100128, 100129, 100130 and 100132), corresponding to the enumerated frame.

Table B1. Decision record for all enumerated records.

ID	Reviewer 1	Reviewer 2	Adjudication	Criterion satisfied at full-text assessment
S1	Include	Include	Not required	Population: application packages. Intervention: detective mechanisms, characterised at family level with feature representations and decision procedures. Comparator: prior surveys, against which the taxonomy and coverage window are positioned. Outcomes: performance figures extracted from surveyed systems. Design: peer-reviewed systematic survey with documented retrieval protocol.
S2	Include	Include	Not required	Population: audio content objects. Intervention: constructive watermarking with full architectural and parameter specification. Comparator: three competing schemes retrained under matched conditions, plus four ablated variants. Outcomes: bit-recovery accuracy, signal-to-noise ratio, area under curve, parameter counts. Design: peer-reviewed primary experimental.
S3	Include	Include	Not required	Population: networked metering devices. Intervention: detective three-phase framework with specified algorithms and hyper-parameters. Comparator: alternative algorithms at each phase plus one external published scheme plus a deliberately degraded configuration. Outcomes: estimation error, accuracy, precision, recall, overhead, latency. Design: peer-reviewed primary experimental.
S4	Include	Include	Not required	Population: broadcast ciphertexts and participant identities. Intervention: constructive cryptographic scheme with complete algorithmic specification. Comparator: three published schemes reimplemented on identical hardware, plus formal hardness assumptions serving as the security baseline. Outcomes: advantage bounds, min-entropy bound, verification outcomes, transmission and computation costs. Design: peer-reviewed primary theoretical with empirical validation.
S5	Include	Include	Not required	Population: cryptographic implementations. Intervention: evaluative attack methodology with complete search-space and parameter specification. Comparator: two conventional search strategies plus three alternative metaheuristics under a matched evaluation budget. Outcomes: guessing entropy, rank statistics with dispersion, execution time, parameter counts. Design: peer-reviewed primary experimental.
E1 (article 100126)	Exclude	Exclude	Not required	Title/abstract screening: Criterion I of Table 1. The article does not present a detective, constructive or evaluative mechanism for authenticity, integrity or provenance; it therefore falls outside the scope of the review as defined in Section 1.2.
E2 (article 100127)	Exclude	Exclude	Not required	Title/abstract screening: Criterion I of Table 1, on the same grounds as the record above.
E3 (article 100130)	Exclude	Exclude	Not required	Title/abstract screening: Criterion I of Table 1, on the same grounds as the records above.
E4 (article 100132)	Exclude	Exclude	Not required	Title/abstract screening: Criterion I of Table 1, on the same grounds as the records above.

Table C1. Rationale for all scores below the maximum.

Study	Domain	Score	rationale
S1	Q4 Data	1	As a survey the study does not itself evaluate on data; it characterises the corpora used by the studies it reviews and identifies their deficiencies, which is appropriate to its genre but constitutes partial rather than full address of the domain.
S1	Q5 Metrics	1	Similarity metrics are defined and their applicability characterised, but they are not applied by the study, and no cost-side metric is systematically extracted across the surveyed corpus.
S1	Q7 Rigour	1	No quantitative synthesis or meta-analysis is attempted across the surveyed studies. The non-mutually-exclusive counting scheme is disclosed transparently, which is creditable, but category frequencies consequently do not support inference about the distribution of approaches.
S2	Q7 Rigour	1	Ablation across four architectural variants and a payload sweep across five sizes are both conducted, which is substantial. No repeated runs or dispersion statistics are reported, so reported margins over comparators cannot be assessed against run-to-run variability in a stochastically trained model.
S2	Q8 Limitations	1	Limitations are confined to the deferral of watermark localisation. Reliance on post-hoc rather than integrated embedding, the absence of adaptive-adversary evaluation, and the encoder-decoder size asymmetry are visible in the reported results but not discussed as constraints.
S3	Q4 Data	1	A single real-world corpus is used, collected for an unrelated purpose, augmented with additional weather descriptors, and supplemented with simulated rather than observed attacks. The construction is described transparently and preserves the physical realism of the measurements, but no independent corpus permits corroboration.
S3	Q6 Baselines	1	Internal comparison across algorithms is thorough at every phase, and the degraded single-neighbour configuration is a well-designed internal control. Only one external published scheme is used as a comparator, and its figures are taken from publication rather than obtained by reimplementing under matched conditions.
S3	Q7 Rigour	1	Overhead and latency are derived analytically with parameters stated, which is a genuine strength. No repeated runs, dispersion statistics or confidence intervals are reported for any classification result.
S4	Q3 Transparency	1	The construction, correctness derivation and security proofs are given in full and an implementation is published. The receiver-authorisation parameter is nevertheless defined inconsistently between the scheme construction and the deployment protocol, the ephemeral group element being present in the former and absent in the latter; since the study's own anonymity argument depends on that element to prevent cross-session linkage, the inconsistency is security-relevant. A typographical corruption also affects one line of the correctness derivation.
S4	Q4 Data	n/a	The contribution is a cryptographic construction whose security is established by reduction and whose efficiency is established by operation counting and timing. It has no evaluation dataset, and none is required by its design. Scoring the domain would measure genre rather than quality, so it is excluded from both numerator and denominator (see §5.3). The desktop and embedded execution timings reported by the study are appraised under Q5 (Methods and Analysis) and Q6 (Comparator Adequacy) rather than Q4, since they concern the transparency of the timing procedure and its comparison against three reimplemented baseline schemes, not the use of a shared benchmark corpus.
S5	Q5 Metrics	1	Metric selection and definition are exemplary, spanning attack effectiveness, dispersion and cost, and including a capacity confound control that is unusual in the architecture-search literature. The score is reduced for reporting consistency: the abstract attributes the best perceptron-space result to one metaheuristic while the corresponding results table attributes that value to a different metaheuristic and records a materially different value for the first. The tabulated value was treated as authoritative under the extraction protocol.

Table E1. PRISMA 2020 checklist with reporting locations.

Item	Checklist item	Reported in
1	Title: identify the report as a systematic review	Title page
2	Abstract: structured summary meeting the PRISMA for Abstracts checklist	Abstract, including registration and funding statements
3	Rationale: describe the rationale in the context of existing knowledge	Section 1.1; Section 1.2
4	Objectives: provide an explicit statement of the objectives or questions addressed	Section 1.3, review questions RQ1 to RQ5
5	Eligibility criteria: specify inclusion and exclusion criteria and grouping for synthesis	Section 2.2; Table 1
6	Information sources: specify all sources and the date each was last searched	Section 2.3; Section 2.4; Table 2; Appendix A
7	Search strategy: present the full search strategy including all filters and limits	Section 2.4; Appendix A, Table A1
8	Selection process: specify methods, number of reviewers and whether they worked independently	Section 2.5; Appendix B
9	Data collection process: specify methods, number of reviewers and any automation used	Section 2.6
10a	Data items: list and define all outcomes sought	Section 2.6; Table 3, fields 19 to 21
10b	Data items: list and define all other variables sought	Section 2.6; Table 3, fields 1 to 18 and 22
11	Study risk of bias assessment: specify methods, tools used and number of reviewers	Section 2.7; Table 4; Appendix C
12	Effect measures: specify the effect measure used for each outcome	Section 2.8; native units retained, no common measure imposed
13a	Synthesis methods: describe processes for deciding eligibility for each synthesis	Section 2.9; grouping by assurance layer and class
13b	Synthesis methods: describe methods of data preparation and conversion	Section 2.6; no conversion or imputation performed
13c	Synthesis methods: describe methods of tabulating and visually displaying results	Section 2.9; Tables 5 to 16; Figures 1 to 5
13d	Synthesis methods: describe methods used to synthesise results	Section 2.9; structured narrative synthesis, five-step procedure
13e	Synthesis methods: describe methods used to explore heterogeneity	Section 2.8; Section 2.9; heterogeneity precluded pooling and is characterised narratively
13f	Synthesis methods: describe any sensitivity analyses conducted	Section 2.9; results reported in Section 4.4, leave-one-out across all themes
14	Reporting bias assessment: describe methods used to assess risk of bias due to missing results	Section 2.10; narrative substitutes specified, funnel-plot methods not applicable
15	Certainty assessment: describe methods used to assess certainty in the body of evidence	Section 2.10; four GRADE-adapted considerations
16a	Study selection: report numbers screened, assessed and included, with a flow diagram	Section 3.1; Figure 1
16b	Study selection: cite studies that appeared to meet the criteria but were excluded, with reasons	Section 3.1; Appendix B; none excluded, and this is stated explicitly
17	Study characteristics: cite each included study and present its characteristics	Section 3.2; Tables 5 and 6
18	Risk of bias in studies: present assessments for each included study	Section 3.3; Figure 3; Table 7; Appendix C
19	Results of individual studies: present summary statistics for each study	Section 3.7; Tables 12 and 13
20a	Results of syntheses: summarise characteristics and risk of bias among contributing studies	Section 3.2; Section 3.3
20b	Results of syntheses: present results of each synthesis	Sections 3.4 to 3.8; Tables 8–14
20c	Results of syntheses: present results of investigations of possible causes of heterogeneity	Section 4.3; heterogeneity attributed to layer and genre
20d	Results of syntheses: present results of sensitivity analyses	Section 4.4
21	Reporting biases: present assessment of risk of bias due to missing results	Section 2.10; Section 5.1; Section 5.4
22	Certainty of evidence: present assessments of certainty for each outcome assessed	Section 4.1; Table 15

Table E1. PRISMA 2020 checklist with reporting locations (continued).

Item	Checklist item	Reported in
23a	Discussion: provide a general interpretation of the results in the context of other evidence	Section 4.1; Section 4.2
23b	Discussion: discuss any limitations of the evidence included	Section 5.4; Section 3.8; Table 14
23c	Discussion: discuss any limitations of the review processes used	Section 5.1; Section 5.2; Section 5.3
23d	Discussion: discuss implications for practice, policy and future research	Section 4.2; Section 6; Table 16; Figure 5
24a	Registration: provide registration information or state that the review was not registered	Section 2.1; Abstract, registration statement
24b	Registration: indicate where the protocol can be accessed or state that none was prepared	Section 2.1; protocol reproduced in Section 2 and Appendices A to C
24c	Registration: describe and explain any amendments to the protocol	Section 5.3; one amendment reported with its effect quantified
25	Support: describe sources of financial or non-financial support and the role of funders	Declarations, Funding and Role of the funding source
26	Competing interests: declare any competing interests of review authors	Declarations, Conflicts of interest
27	Availability of data, code and other materials: report which materials are publicly available and where	Declarations, Data availability; Appendices A to C