SSR Journal of Artificial Intelligence (SSRJAI)

OPEN ACCESS

Homepage: https://ssrpublisher.com/ssrjai/

Volume 3, Issue 7, 2026 ISSN: 3049-0413

Generative Artificial Intelligence for Synthetic Network Traffic Generation in Intrusion Detection: A Systematic Review and Future Directions

G.I.O. Aimufua[1] and M.S. Bandiya[2]

[1]Centre for Cyberspace Studies, Nasarawa State University, Keffi (NSUK), Nasarawa State, Nigeria

[2]Department of Cybersecurity, Centre for Cyberspace Studies, Nasarawa State University, Keffi (NSUK), Nasarawa State, Nigeria

Received: *05.06.2026* | ***Accepted:*** *13.07.2026* | ***Published:*** *15.07.2026*

***Corresponding author:** M.S. Bandiya

DOI: *10.5281/zenodo.21367660*

Abstract **Original Research**

Machine learning is central to modern network intrusion detection, yet its performance is constrained by two chronic data pathologies: severe class imbalance between abundant benign traffic and rare attack flows, and the scarcity of labelled attack data arising from privacy, legal, and operational constraints on capturing real malicious traffic. Generative artificial intelligence, encompassing generative adversarial networks, variational autoencoders, and diffusion models, has been advanced as a remedy that synthesises realistic attack traffic to rebalance datasets and enrich training corpora. Despite a rapidly expanding body of primary studies, the literature lacks a consolidated assessment of three interdependent concerns: the fidelity of synthetic traffic, its downstream utility for detection, and the security risks that synthetic data may itself introduce. This article reports a systematic review, conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guideline, of generative approaches to synthetic network traffic generation for intrusion detection published between 2020 and 2026, together with the protocol and future research directions. Framed by generative learning theory and a seven-type taxonomy of research gaps, the review identifies the methodological, empirical, and knowledge gaps the field exhibits, and advances a three-dimensional framework separating the fidelity, utility, and security dimensions of synthetic-traffic quality. The methodology specifies the search strategy across five databases, the eligibility criteria, the screening and data-extraction procedures, and the quality-appraisal instruments. The contribution is threefold: a reproducible review protocol, a three-dimensional evaluation framework, and future directions foregrounding the under-examined security implications of synthetic data.

Keywords: Generative artificial intelligence, Synthetic network traffic, Intrusion detection, Generative adversarial networks, Diffusion models, Class imbalance, Systematic review.

1. Introduction

Network intrusion detection systems form a foundational layer of contemporary cyber defence, tasked with distinguishing malicious activity from the immense volume of benign traffic that traverses modern enterprise, cloud, and Internet of Things networks [1]. Over the past decade, the field has shifted decisively from signature-based detection, which recognises only

previously catalogued attack patterns, towards machine-learning and deep-learning detectors capable of learning discriminative representations of normal and anomalous behaviour directly from data. This shift has yielded detectors that report high accuracies on widely used benchmarks, yet the headline figures conceal a structural fragility rooted not in the detection algorithms themselves but in the data on which they are trained [2].

Two data pathologies recur across the empirical literature. The first is class imbalance. In realistic traffic captures, benign flows outnumber attack flows by several orders of magnitude, and certain attack categories are represented by only a handful of instances. In the widely used CIC-IDS2017 benchmark, for example, a single attack class is represented by 11 flow records against more than two million benign records, a ratio that systematically biases classifiers towards the majority class and depresses the detection of precisely the rare, high-consequence attacks that matter most [3]. The second pathology is the scarcity of labelled attack data. Capturing authentic malicious traffic at scale is impeded by privacy regulation, the legal sensitivity of intrusion records, the cost of expert labelling, and the operational difficulty of provoking diverse, well-documented attacks in controlled environments [4]. The two pathologies compound one another: the classes that are scarcest are also those most likely to be under-labelled, leaving detectors blind to the long tail of the threat distribution.

Generative artificial intelligence has emerged as the leading data-centric response to these pathologies. Rather than redesigning the detector, the generative paradigm augments the training corpus with synthetic traffic that mimics the statistical and behavioural properties of real attacks, thereby rebalancing classes and enriching scarce categories. Three families of generative models dominate the field. Generative adversarial networks pit a generator against a discriminator in a minimax game [5], and their conditional, Wasserstein, and tabular variants have been applied extensively to oversample minority attack classes. Variational autoencoders encode traffic into a structured latent space from which new samples can be decoded, offering training stability at some cost to sample sharpness [6]. Most recently, diffusion models generate data by progressively denoising a noise prior [7], and have been adapted to tabular and time-series traffic. The detailed application of each family to intrusion detection is examined in Section 3.

Notwithstanding this proliferation, the evidence base is fragmented along three axes that no existing synthesis adequately reconciles. The first is fidelity: whether synthetic traffic faithfully reproduces the marginal distributions, feature dependencies, and protocol semantics of real traffic, and by which metrics fidelity should be judged. The second is utility: whether augmentation with synthetic data measurably improves downstream detection performance, and whether reported gains generalise beyond the benchmark on which they were obtained. The third, conspicuously under-examined, is security: whether the synthetic data and the generative pipeline themselves introduce new risks, including the leakage of sensitive records through membership-inference attacks (MIA) and the dual-use potential of generators to craft evasive adversarial traffic [8,9]. The same architecture that balances a defender's dataset can equip an adversary to synthesise traffic engineered to evade detection [10].

This article reports a systematic review, conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 guideline [11], of generative approaches to synthetic network traffic generation for intrusion detection published between 2020 and 2026, together with its protocol and a set of future research directions. The review is guided by four questions: (i) which generative architectures have been applied to synthetic network traffic generation for intrusion detection, and how have they evolved; (ii) which datasets and metrics have been used to evaluate synthetic-traffic fidelity and downstream detection utility; (iii) what evidence exists regarding the security and privacy risks introduced by synthetic data and generative pipelines; and (iv) what

methodological and empirical gaps remain to be addressed.

The contribution is threefold, and is positioned deliberately against the closest prior synthesis. A recent PRISMA-compliant review of generative adversarial networks for cybersecurity defence [1] catalogues architectures and organises them by defensive function, architecture, domain, and threat model. The present review departs from that descriptive, architecture-centred orientation in three ways. First, it delivers a reproducible review protocol focused specifically on synthetic network-traffic generation for intrusion detection. Second, and centrally, it advances an evaluation framework that separates the fidelity, utility, and security dimensions of synthetic-traffic quality and treats them as distinct, only weakly correlated axes; whereas prior taxonomies classify methods by what they are, this framework classifies evidence by how synthetic-traffic quality should be judged. Third, it frames a research agenda that elevates the neglected security implications of synthetic data from a peripheral observation to a first-order organising concern. The remainder of the article is organised as follows. Section 2 develops the theoretical and conceptual framework and applies a taxonomy of research gaps to the field. Section 3 reviews the prior literature. Section 4 details the systematic review methodology.

2. Theoretical and conceptual framework

A systematic review of a methodologically heterogeneous field requires an explicit theoretical scaffold that both motivates the synthesis and disciplines the interpretation of its findings. This section establishes that scaffold in four movements. It first articulates the generative learning theory underpinning the three model families under review. It then adopts and applies a taxonomy of research gaps as the analytical lens through which the literature is interrogated, identifying the specific gaps this study addresses. It finally integrates these into a conceptual framework linking the data pathologies of intrusion detection to the generative remedies and to the evaluation dimensions by which those remedies must be judged.

2.1. Generative learning theory and the three model families

Generative modelling seeks to learn a representation of the probability distribution that produced a body of observed data, such that new samples can be drawn that are statistically indistinguishable from real observations. Three theoretical formulations of this objective dominate the field, each embodying a distinct inductive bias and a distinct set of trade-offs between fidelity, diversity, and training stability.

The adversarial formulation, realised in generative adversarial networks, frames generation as a two-player minimax game in which a generator learns to produce samples that a discriminator cannot distinguish from real data [5]. At the game's equilibrium the generated distribution converges to the data distribution without requiring an explicit likelihood. Its theoretical liability is instability: the adversarial objective is prone to non-convergence and mode collapse, in which the generator captures only a subset of the data's modes and thereby fails to reproduce the very diversity that minority-class augmentation demands [12]. Conditional and Wasserstein reformulations were introduced to stabilise training and to permit class-conditional generation.

The variational formulation, realised in variational autoencoders, frames generation as the maximisation of a lower bound on the data likelihood, learning an encoder that maps data into a structured latent distribution and a decoder that reconstructs data from latent samples [6]. Its strength is a stable, likelihood-grounded objective and an interpretable latent space; its characteristic weakness is a tendency to produce smoothed, less sharply defined samples, which in the traffic setting can blur the fine-grained feature interactions that distinguish attack from benign flows [13]. Hybrid variational-adversarial architectures combine the stability of the former with the sharpness of the latter [14].

The diffusion formulation, realised in denoising diffusion probabilistic models, frames generation as the learned reversal of a gradual noising process [7]. Its advantages are stable optimisation and strong mode coverage, which directly address the mode-collapse liability of adversarial training; its principal costs are

computational and the need to handle the mixed continuous and categorical structure of tabular traffic features. Latent-space diffusion, in which a learned autoencoder first compresses heterogeneous features into a uniform latent representation, mitigates both costs [15]. The progression from adversarial to variational to diffusion formulations constitutes the dominant arc of architectural evolution that the review traces.

2.2. A taxonomy of research gaps as analytical lens

To interrogate the literature systematically rather than impressionistically, the review adopts the seven-type taxonomy of research gaps, which distinguishes evidence, knowledge, practical-knowledge, methodological, empirical, theoretical, and population gaps and offers a disciplined vocabulary for locating where prior research is incomplete, contradictory, or absent. An evidence gap arises where findings across studies are mutually contradictory at an abstract level; a knowledge gap where desired findings simply do not exist; a practical-knowledge gap where professional practice diverges from research findings; a methodological gap where prevailing research methods are insufficient or insufficiently varied to generate trustworthy insight; an empirical gap where propositions remain unverified by direct evaluation; a theoretical gap where the explanatory theory is dated or absent; and a population gap where particular sub-populations remain under-represented.

2.3. Application of the taxonomy: gaps addressed by this review

Applying the taxonomy to the present field locates three specific gaps, which this review is designed to address and which are written up below following a structured form for gap identification.

The review identifies a methodological gap in the prior research concerning the evaluation of synthetic network traffic. Previous research has addressed several aspects of generative augmentation for intrusion detection: adversarial oversampling of minority classes [16,17], tabular and hybrid generation [14,18], and diffusion-based synthesis [15,19]. However, the prior research has not established a consolidated evaluation methodology: studies are conducted on disparate benchmarks with incommensurable metrics, so that synthetic-traffic quality cannot be compared across studies. A variation and standardisation of evaluation methods is therefore necessary to generate trustworthy, cumulative insight.

An empirical gap also remains in the prior research. Reported improvements in downstream detection performance are seldom evaluated for generalisation beyond the originating benchmark, and the recurring design flaws of the benchmarks themselves cast doubt on whether reported gains transfer to operational settings [2]. No body of work to date has directly and systematically verified whether the utility gains attributed to synthetic augmentation hold across datasets and deployment conditions. An empirical synthesis of these issues is important because the credibility of the entire augmentation paradigm rests on the transferability of its reported gains.

The review identifies a knowledge gap concerning the security of synthetic data. The prior augmentation research did not adequately address the subject of the risks introduced by the synthetic data and the generative pipeline themselves. Adjacent literatures show that synthetic data does not automatically confer anonymity [8,20], that generative models are subject to membership-inference [9,21], and that the same generators can be turned to adversarial evasion [10,22], yet these findings remain largely unconnected to intrusion-detection augmentation. This subject should be explored further to provide an understanding of why security has not been treated as a first-order concern in the field.

In response to these three gaps, this review seeks to extend the evidence base by, first, constructing and applying a three-dimensional framework that addresses the methodological gap in evaluating synthetic-traffic quality; second, interrogating the empirical gap by recording whether reported utility gains were tested for generalisation; and third, addressing

the knowledge gap by treating the security implications of synthetic data as a deliberate object of synthesis rather than an incidental finding.

2.4. The conceptual framework

The conceptual framework integrates the foregoing into a single explanatory structure linking problem, remedy, and evaluation. It posits a causal chain in which the data pathologies of intrusion detection, namely class imbalance and labelled-data scarcity, motivate the application of generative remedies drawn from the three model families, whose synthetic outputs must then be appraised along three evaluation dimensions before any claim of improvement can be sustained. Fidelity concerns the degree to which synthetic traffic reproduces the statistical and behavioural properties of real traffic. Utility concerns the degree to which augmentation improves downstream detection performance. Security concerns the degree to which the synthetic data and generative pipeline introduce new risk, encompassing privacy leakage and the dual-use potential for adversarial evasion [8,20].

The framework's central proposition is that fidelity, utility, and security are distinct and only weakly correlated, such that a method may excel on one dimension while failing on another. High-fidelity synthesis, for instance, may heighten privacy risk precisely because faithfully reproduced records can disclose information about the systems whose traffic was modelled [20]. The framework therefore mandates that synthetic-traffic quality be assessed multi-dimensionally, and it supplies the structure against which the review's extracted evidence is organised. The framework is depicted in Fig. 1.

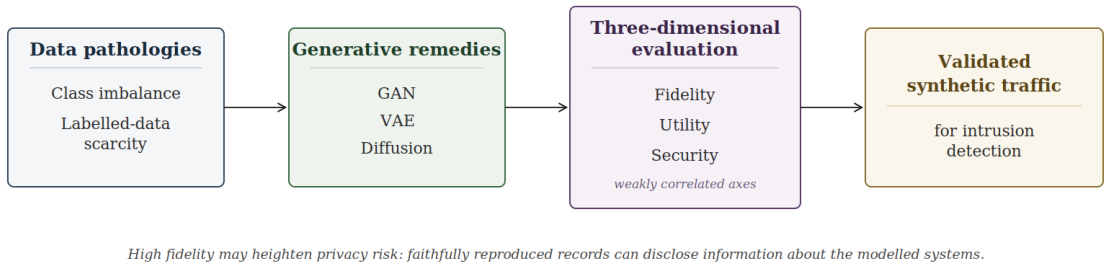

Fig. 1. Conceptual framework linking the data pathologies of intrusion detection to generative remedies and the three-dimensional evaluation of synthetic-traffic quality.

Note. A high-resolution vector version of Fig. 1 accompanies the manuscript as a separate artwork file, per journal submission requirements.

3. Literature review

This section reviews the prior research the systematic review interrogates, organised thematically so that the evolution of architectures, the conduct of evaluation, and the treatment of security can each be appraised in turn.

3.1. Adversarial approaches

Generative adversarial networks are the most heavily represented architecture family in the intrusion-detection literature. Conditional and modified conditional architectures have been advanced specifically for class-imbalance problems, reporting improved minority-class

detection on standard benchmarks [16], and imbalanced adversarial designs coupling a generator with downstream sequence classifiers report high multi-class accuracy after augmentation [17]. Tabular adversarial networks, which respect the mixed continuous and categorical structure of flow features, synthesise minority attack samples without distorting the statistical properties of real traffic [18]. Multi-generator and self-attention designs improve mode coverage and capture long-range dependencies among features [12], and hybrid edge-deployed schemes extend the approach to Internet of Things settings [23]. The recurring finding is that adversarial augmentation raises minority-class recall and overall accuracy relative to unaugmented or classically oversampled baselines; the recurring limitation is methodological, in that evaluations use disparate benchmarks and metrics and the persistent threat of mode collapse means reported diversity is seldom rigorously verified.

3.2. Variational and hybrid approaches

Variational autoencoders occupy a smaller but methodologically important segment, valued for training stability and interpretable latent representations. Anomaly-based detectors built on variational reconstruction error exploit the premise that attack traffic reconstructs poorly under a model trained predominantly on benign traffic [24]. Conditional variational designs, including log-cosh formulations, have been used both to detect intrusions and to generate minority samples [13]. Hybrid conditional Wasserstein variational-adversarial architectures coupled with convolutional classifiers address class imbalance with competitive multi-class performance [14]. The literature converges on the assessment that variational methods offer stability at the cost of sample sharpness, and that hybridisation is the dominant strategy for reconciling the trade-off.

3.3. Diffusion and emerging paradigms

Diffusion models are the newest and fastest-growing paradigm. Time-series imaging approaches transform traffic into image-like representations and apply diffusion to synthesise network traffic, reporting substantial fidelity and downstream gains over adversarial baselines [19]. Tabular diffusion models, including score-based latent-space and mixed-type designs, are benchmarked against adversarial and variational baselines and report superior fidelity and downstream utility [15,25], and multimodal generative frameworks that jointly synthesise tabular and image-transformed representations report further gains in fidelity, diversity, and detection performance [26]. A cautionary strand demonstrates the dual-use character of diffusion: diffusion-generated adversarial traffic has been shown to deceive deep-learning attack detectors [22].

3.4. Datasets and evaluation practice

For intrusion-detection evaluation the field rests heavily on a small set of public network benchmarks, principally NSL-KDD, CIC-IDS2017, CIC-IDS2018, UNSW-NB15, and a growing family of Internet of Things datasets such as TON_IoT and CICIoT2023 [3,27]; a separate set of image corpora, profiled alongside these in Section 5.2, is drawn upon only where the privacy leakage of diffusion models has been most rigorously quantified. These benchmarks are indispensable yet problematic: systematic analyses identify recurring design flaws, including limited feature diversity, high feature interdependence, ambiguous ground truth, collapsed traffic distributions, and labelling inaccuracies, which correlate with poor generalisation to operational environments [2]. Evaluation metrics are correspondingly heterogeneous, spanning statistical distance, fidelity scores, precision-recall-density-coverage, machine-learning efficacy, and detectability [26]. A dedicated fidelity-evaluation framework argues that accuracy-centred evaluation is insufficient and that fidelity must be appraised in its own right [28]. The absence of a shared, security-inclusive evaluation protocol is the methodological gap this review's framework is designed to close.

3.5. Security and privacy implications

The security dimension is the least developed and the most consequential. Synthetic data are

frequently presumed privacy-preserving, yet empirical work demonstrates that high-fidelity synthesis can leak information about training records and does not automatically confer anonymity [8,20]. Membership-inference attacks succeed against both machine-learning models and diffusion models, implying that generative intrusion-detection pipelines may expose sensitive operational traffic [9,21]. Differentially private synthesis built directly on the differential-privacy framework reports favourable fidelity-privacy trade-offs in downstream anomaly detection [29]. On the dual-use side, the generative architectures that balance a defender's dataset can synthesise traffic engineered to evade detection [10,22], and recent surveys recognise this defensive-offensive co-evolution [1]. Despite the gravity of these risks, the augmentation literature seldom reports privacy audits or evasion-robustness assessments.

3.6. Prior surveys and the identified gaps

Prior surveys advance the field descriptively. A PRISMA-compliant review of generative adversarial networks for cybersecurity defence retained 185 studies and developed a four-dimensional taxonomy spanning defensive function, architecture, domain, and threat model [1], and reviews of generative models for synthetic attack data survey the breadth of model families [4]. Yet three gaps persist, as set out in Section 2.3: no existing synthesis integrates fidelity, utility, and security into a single evaluation framework (methodological gap); the generalisation of reported utility gains remains empirically unverified (empirical gap); and the security and privacy risks of synthetic data are treated peripherally (knowledge gap). The present review is designed to address all three.

4. Methodology

This review adopts a systematic, protocol-driven methodology to ensure that the synthesis is transparent, reproducible, and resistant to selection bias, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 guideline [11]. The methodology is reported across seven components: research philosophy and design, the review protocol and questions, the search strategy, the eligibility criteria, the study-selection process, the data-extraction scheme, and the quality-appraisal and synthesis procedures.

4.1. Research philosophy and design

The review is situated within a post-positivist research philosophy, which holds that an objective reality exists and can be approximated through systematic, replicable inquiry while acknowledging that all measurement is fallible. Consistent with this stance, the review adopts a systematic literature review design augmented by a structured, framework-guided narrative synthesis. A purely meta-analytic design is deliberately not adopted, because the heterogeneity of architectures, datasets, and metrics renders the effect sizes statistically incommensurable; pooling them would manufacture a spurious precision. Framework-guided narrative synthesis, organised by the fidelity–utility–security framework of Section 2.4, is the design best matched to a field whose principal pathology is incommensurability.

4.2. Review protocol and research questions

A review protocol was defined in advance of screening and specifies the research questions, the search strategy, the eligibility criteria, the screening procedure, the extraction scheme, and the appraisal instruments; it is available from the authors on request. The review is guided by the four research questions stated in Section 1, and each included study will be interrogated against all four, with the framework dimensions providing the coding categories against which evidence is mapped.

4.3. Search strategy

A comprehensive search is designed to maximise sensitivity while preserving precision. Five bibliographic databases are searched, namely Scopus, IEEE Xplore, the ACM Digital Library, ScienceDirect, and SpringerLink, supplemented by forward and backward citation chasing and by a targeted arXiv search to capture

recent preprints. The search string combines three concept blocks with Boolean operators: a generative-methods block (generative adversarial network, GAN, variational autoencoder, VAE, diffusion model, generative artificial intelligence), a data-objective block (synthetic traffic, synthetic data, data augmentation, traffic generation, class imbalance), and an application-domain block (intrusion detection, network security, anomaly detection), combined as the conjunction of three disjunctions. The search is restricted to January 2020 through the 2026 search date, with foundational theoretical references predating the window retained where they define the architectures under review. The complete database-specific search strings are provided in Appendix A.

4.4. Eligibility criteria

Inclusion and exclusion criteria are defined a priori. A study is eligible if it applies one or more generative models, namely a generative adversarial network, a variational autoencoder, a diffusion model, or a documented hybrid, to the synthesis of network traffic or flow data; evaluates the synthetic data within an intrusion- or anomaly-detection context; reports an empirical evaluation with extractable methods and results; is published in English; and appears between January 2020 and the search date, foundational architectural references excepted. To resolve the boundary between general tabular synthesis and traffic synthesis, a general mixed-type tabular generator is included only where it is evaluated on a traffic or flow dataset, or where its findings are explicitly transferred to intrusion detection; general tabular-synthesis studies with no traffic-specific evaluation are treated as background and excluded from the synthesised corpus. A study is excluded if it uses only classical, non-generative oversampling; generates data for a non-network domain without transferable findings; is a non-substantive non-peer-reviewed source; or duplicates an included study. These criteria are summarised in Table 1.

Table 1. Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
Applies a GAN, VAE, diffusion model, or documented hybrid to network-traffic synthesis	Uses only classical non-generative oversampling (e.g., SMOTE) with no generative model
Evaluates synthetic data within an intrusion- or anomaly-detection context	Generates data for a non-network domain with no transferable findings
General tabular generator included only if evaluated on a traffic/flow dataset or explicitly transferred to intrusion detection	General mixed-type tabular synthesis with no traffic-specific evaluation (background only)
Reports an empirical evaluation with extractable methods and results; English; 2020–2026 (foundational refs exempt)	Non-substantive non-peer-reviewed source, or duplicate of an included study

4.5. Study selection

Records retrieved from all sources are imported into a reference-management system and de-duplicated. Screening proceeds in two stages: titles and abstracts are screened against the eligibility criteria, then full texts of the remaining records are assessed in detail, with reasons for exclusion recorded for every full-text rejection. To safeguard reliability, records are screened against the eligibility criteria in a

structured, documented procedure, and borderline cases are resolved by re-examination against the full text and the a priori criteria. The number of records identified, screened, excluded, and included at each stage will be reported in a PRISMA 2020 flow diagram, the template for which is shown in Fig. 2.

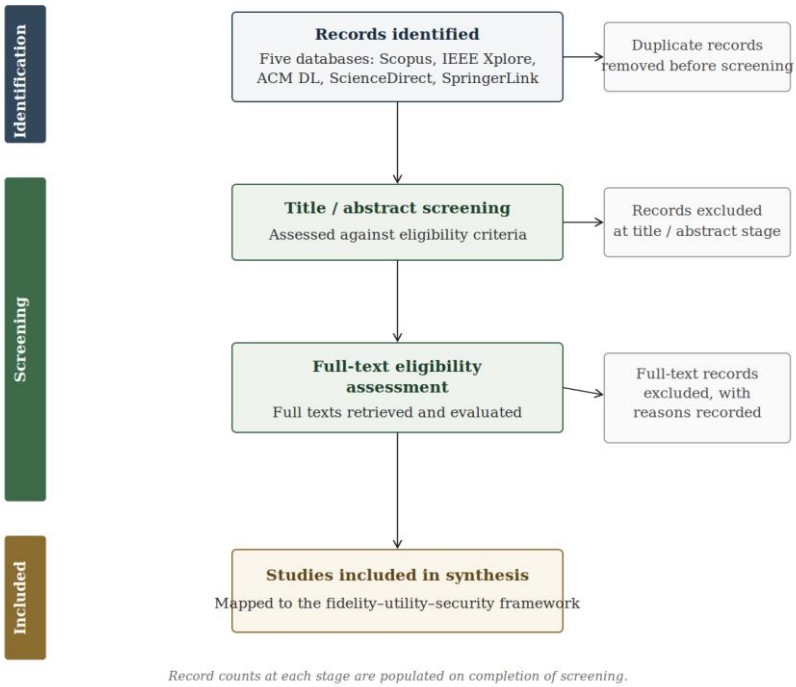

Fig. 2. PRISMA 2020 flow-diagram template for the study-selection process.

Note. The populated diagram with record counts at each stage is produced on completion of screening and accompanies the manuscript as a separate artwork file.

4.6. Data extraction

A standardised data-extraction form is developed and piloted on a sample of included studies before full application. For each included study the form captures bibliographic metadata; the generative architecture and any hybridisation; the datasets used; the preprocessing and feature representation; the evaluation metrics reported, classified under the fidelity, utility, and security dimensions; the principal quantitative results; the reported limitations; and any treatment of privacy or adversarial risk. The classification of every reported metric under one of the three framework dimensions is the analytical core of the extraction, because it renders heterogeneous studies comparable along common axes. Extraction is performed against a piloted standardised form, with entries reconciled against the source study to reduce transcription error. The extraction scheme is summarised in Table 2.

Table 2. Data-extraction scheme mapped to the three-dimensional evaluation framework.

Extraction category	Items captured
Study metadata	Authors, year, venue, study type, application setting (enterprise, cloud, IoT)
Generative method	Architecture family (GAN, VAE, diffusion, hybrid), specific variant, conditioning strategy
Data and preprocessing	Datasets used, feature representation, normalisation, encoding, train/test protocol
Fidelity dimension	Statistical distance, fidelity scores, precision-recall-density-coverage, detectability
Utility dimension	Machine-learning efficacy, minority-class recall and F1, accuracy gains, generalisation tests
Security dimension	Privacy audits, membership-inference resistance, differential-privacy guarantees, evasion robustness, dual-use discussion
Limitations	Author-reported limitations and threats to validity

4.7. Quality appraisal and synthesis

The methodological quality of each included study is appraised using a structured rubric that adapts the Joanna Briggs Institute critical-appraisal approach and recognised machine-learning reproducibility criteria to the computational intrusion-detection setting. The rubric assesses objective clarity, dataset appropriateness and provenance, train-test rigour and guard against leakage, metric comprehensiveness relative to the three framework dimensions, reproducibility including code and data availability, generalisation testing, and candour of limitation and risk disclosure. Each criterion is scored on a three-point ordinal scale, and the resulting profile weights the confidence attached to each study's findings during synthesis and supports a structured risk-of-bias assessment across the corpus. The appraisal criteria are summarised in Table 3.

Table 3. Quality-appraisal rubric (adapted from JBI critical appraisal and ML-reproducibility criteria; scored 0 = not met, 1 = partially met, 2 = fully met).

Appraisal criterion	Question addressed
Objective clarity	Is the research objective and the synthetic-data role stated explicitly?
Dataset appropriateness	Are datasets suitable, current, and described, with provenance and known flaws acknowledged?
Train-test rigour	Is the train-test separation sound and protected against leakage between real and synthetic partitions?
Metric comprehensiveness	Do reported metrics cover fidelity, utility, and security rather than accuracy alone?
Reproducibility	Are code, hyperparameters, and data availability sufficient to reproduce the procedure?
Generalisation testing	Are gains tested beyond the originating benchmark (cross-dataset or operational validation)?
Limitations and risk disclosure	Are limitations, privacy implications, and dual-use risks candidly disclosed?

Synthesis will proceed by the framework-guided narrative method. Extracted evidence is aggregated within each of the three evaluation dimensions and across the three architecture families, so that the synthesis can report not only what each method achieves in isolation but how the dimensions interact, where the literature is contradictory, and where it is silent. Particular attention is directed to the central proposition that fidelity, utility, and security are only weakly correlated. As a review of published literature, the study involves no human participants and requires no ethical approval; it nonetheless adheres to the ethics of evidence synthesis by reporting the protocol transparently, recording exclusion reasons, and representing primary studies faithfully, and it treats the dual-use character of the technologies under examination as a deliberate object of synthesis.

5. Results

The evidence assembled by this review is organised along the three axes of the fidelity–utility–security framework advanced in Section 2.4. This section first maps the reviewed primary studies to the three pillars, then reports the synthesised findings in that order: the statistical fidelity of synthetic traffic, the downstream utility of generative augmentation for intrusion detection, and the security and privacy risks that synthetic-data pipelines themselves introduce. Where the strongest privacy evidence derives from image-based generative models, it is used to characterise the diffusion mechanism and its transfer to network data is treated explicitly. Because the corpus spans generative adversarial networks, variational autoencoders, and the newer diffusion family, the results are presented so that findings which recur across architectures are distinguished from those confined to a single model class [54,56]. A preliminary characterisation of the datasets and attack taxonomies that underpin the primary studies is given first, since the interpretation of every downstream metric is conditioned on the data on which it was obtained. All quantitative values reported in this section are secondary data extracted from the primary studies included in the review rather than measurements produced by the authors; each is attributed to its originating study, and the tables reproduce figures as reported in those sources.

5.1. Overview of the reviewed primary studies

This review synthesises evidence from a corpus of primary studies spanning the three generative families and the three evaluation dimensions. Table 4 presents the literature mapping that anchors the synthesis, recording, for each core study, the generative model, the research focus, the principal reported result, and the gap the study leaves open. The mapping is organised by the fidelity, utility, and security pillars so that convergent and divergent findings can be read across studies, and its entries frame the dimension-by-dimension analysis that follows. Because network-traffic studies rarely share a benchmark or metric, the reported figures are not pooled; they are read comparatively within each pillar.

Table 4. Literature mapping of the reviewed primary studies by research pillar, model, focus, key result, and identified gap.

Study (ref.)	Model	Core focus	Key reported result	Identified gap
Sivaroopan et al. [19]	NetDiffus	1D traffic converted to 2D GASF images to avoid GAN mode collapse	66.4% higher fidelity than SOTA GANs; FID < 20	Generates traffic features only, not network metadata
Shi et al. [15]	TabDiff	Joint continuous-time diffusion for mixed-type tabular data	22.5% improvement in pair-wise column correlation	High encoding overhead not explicitly addressed
Zhang et al. [25]	TabSyn	Diffusion within a VAE-crafted latent space	86% reduction in column-wise distribution error	Depends on initial VAE training quality
Rahman et al. [23]	SYN-GAN	Training NIDS on 100% synthetic data for IoT security	100% accuracy on BoT-IoT; 90% on UNSW-NB15	Poor adaptation to unseen emerging threats
Allagi et al. [18]	CTGAN-CNN	Conditional tabular GAN balancing of benchmark data	NSL-KDD balanced 50/50; 95.47% detection accuracy	Lacks advanced validation (duplicate detection, t-SNE)
Rao & Babu [17]	IGAN	Discovery of minority attack classes in imbalanced data	>98% accuracy with LeNet-5 + LSTM ensemble	Compute limits deeper networks with more residual blocks
He et al. [14]	CWVAEGAN-IDS	Non-Gaussian continuous features via variational Gaussian mixtures	90.11% accuracy on KDDTEST+; outperforms standard GANs	Learning degrades when training sets are excessively large
Sun et al. [29]	NetDPSyn	Network-trace synthesis under differential privacy	0.90 Spearman correlation on TON_IoT; matches raw utility	Coarse temporal representation misses protocol dependencies
Meeus et al. [20]	Distance-based ID	Principled identification of vulnerable synthetic records	+7.2 percentage points over ad-hoc methods	Ordinal features and higher-order distances unexplored
Stadler et al. [8]	Black-box evaluation	Synthetic data vs. traditional sanitisation for linkability	Often no better privacy-utility trade-off than sanitisation	Outlier records remain vulnerable to linkage attacks
Kumar et al. [22]	NetDiffuser	Natural adversarial examples to deceive NIDS detectors	Detector AUC-ROC reduced by up to 0.534	Ensuring functional validity while preserving stealth

SOTA = state of the art; FID = Fréchet Inception Distance; NIDS = network intrusion detection system; DP = differential privacy. All figures are reported as stated in the cited primary studies.

5.2. Characterisation of datasets and evaluation substrates

The included studies evaluate generative and detection methods across a heterogeneous set of substrates that fall into three broad families: tabular network-flow datasets used for intrusion detection, flow-based captures used as sources for synthetic-traffic generation, and image corpora repurposed for the controlled study of membership-inference vulnerability. This heterogeneity is itself a finding: the field has no canonical benchmark on which fidelity, utility, and privacy can be jointly measured, which is the structural incommensurability that motivates a framework-guided rather than meta-analytic synthesis. Table 5 profiles the principal datasets recovered from the corpus, together with their data type, primary use case, and salient statistics.

Table 5. Profiles of the principal datasets used across the reviewed studies for synthetic-traffic generation, intrusion detection, and membership-inference evaluation.

Dataset	Data type	Primary use case	Key statistics
CIFAR-10	Image	MIA evaluation	50,000 training images; 25,000 member / 25,000 hold-out (50% split).
CIFAR-100	Image	MIA evaluation	50,000 training images; 25,000 member / 25,000 hold-out (50% split).
NSL-KDD	Tabular	Intrusion detection	41 features (9 discrete, 32 continuous); DoS, Probe, U2R, and R2L classes.
STL10-U	Image	MIA evaluation	100,000 unlabelled images from STL10; used for DDPM vulnerability testing.
Tiny-ImageNet	Image	MIA evaluation	100,000 training images; 50,000 member / 50,000 hold-out (50% split).
Pokemon	Text-to-image	Synthetic generation	Fine-tuning set for LDMs; highly centralised, uniform stylistic image/text pairs.
COCO2017-val	Text-to-image	MIA evaluation	2,500 samples used for hold-out sets in large-scale Stable Diffusion benchmarks.
Laion-aesthetic-5plus	Text-to-image	Training / MIA	Aesthetic scores > 5; 2,500 samples used for member sets in Stable Diffusion testing.
CIDDS-001	Flow-based	Synthetic traffic	Source for WGAN-based generation; IP, ports, protocols, and transport features.
CMU-CERT	Tabular	Insider-threat detection	Discrete and continuous columns; used for CTGAN-based class balancing.

MIA = membership-inference attack; LDM = latent diffusion model; DDPM = denoising diffusion probabilistic model; WGAN = Wasserstein generative adversarial network.

The tabular intrusion-detection datasets, exemplified by NSL-KDD with its 41 features partitioned across discrete and continuous columns and its four canonical attack families, remain the workhorses for measuring the downstream utility of class-balancing generators such as conditional tabular GANs [33–35]. Flow-based captures such as CIDDS-001 supply the transport-layer structure that Wasserstein-penalised generators must reproduce faithfully

[31,36], while the CMU-CERT corpus anchors the insider-threat studies in which CTGAN is used to rebalance rare malicious behaviour [46]. The image corpora, though drawn from computer vision rather than networking, are retained in the synthesis because they furnish the controlled conditions under which the privacy leakage of diffusion models has been most rigorously quantified; the network-security relevance of those measurements is developed in Section 5.5.

5.3. Dimension I: fidelity of synthetic traffic and synthetic samples

Fidelity concerns the degree to which synthetic samples reproduce the statistical and behavioural properties of real data. Across the reviewed generative families, fidelity is assessed with three complementary instruments: distributional distance measured by the Fréchet Inception Distance (FID), information-loss norms comparing synthetic and real feature moments, and diversity scores that guard against mode collapse. The evidence indicates that the most recent diffusion-based generators attain fidelity so high that synthetic and real samples become difficult to separate on distributional grounds alone, a property that is desirable for augmentation but, as Section 5.5 shows, simultaneously enlarges the privacy attack surface [44,45].

Table 6 reports the Fréchet Inception Distance obtained for a representative diffusion model on member and hold-out partitions, as extracted from the diffusion membership-inference literature [9]. The near-identical scores of 9.66 and 9.85 demonstrate that the generator achieves high synthesis quality without exhibiting a pronounced bias toward the training distribution: from the vantage point of a coarse distributional metric, the model does not obviously overfit. This absence of memorisation at the aggregate level is precisely why finer, step-wise instruments are required to expose the residual leakage that aggregate metrics conceal.

Table 6. Fréchet Inception Distance (synthetic vs. real) for a diffusion model evaluated on member and hold-out partitions; lower is better.

Sample type	FID score (synthetic vs. real)
Member samples	9.66
Hold-out samples	9.85

A more discriminating fidelity instrument is the step-wise reconstruction error, or t-error, which measures how accurately a deterministic reverse–denoise cycle reconstructs a sample under the learned model. Table 7 summarises its behaviour for member and hold-out sets on CIFAR-10 and Tiny-ImageNet, as reported in the source study [9], which presents the t-error in relative terms with the member set normalised to unity. On both datasets the member sets reconstruct with markedly lower error than the hold-out sets. The divergence is systematic rather than incidental: the hold-out error is both larger and several times more variable than the member error, confirming that the model reconstructs training members more tightly and more consistently than unseen samples. This asymmetry is the fidelity signature that doubles as a privacy vulnerability, and it is the quantity exploited by the strongest attacks reported in Section 5.5.

Table 7. Relative step-wise reconstruction error (t-error) for member and hold-out sets, with the member-set value normalised to 1.00, as reported qualitatively in the source study [9].

Dataset	Set type	Relative t-error (member = 1.00)	Relative variance
CIFAR-10	Member set	1.00	1.0
CIFAR-10	Hold-out set	higher	several × member
Tiny-ImageNet	Member set	1.00	1.0
Tiny-ImageNet	Hold-out set	higher	several × member

Values are relative, with each dataset's member-set t-error normalised to 1.00; the source study reports the member/hold-out asymmetry qualitatively rather than as absolute magnitudes. Lower relative t-error on member sets indicates tighter reconstruction of training data by the learned model.

For generative adversarial networks the fidelity picture is assessed with a different family of instruments, summarised in Table 8 [32]. The Inception Score is used to gauge the diversity and class-conditional quality of samples from architectures in the StackGAN lineage; multi-scale statistical similarity checks whether the local structure of progressively grown images matches the training set across resolutions; and an information-loss norm quantifies the discrepancy between the means and standard deviations of synthetic and real tabular features. Taken together, these instruments confirm that adversarial generators can achieve strong perceptual and statistical fidelity, but that fidelity is fragile: mode collapse and unstable training remain recurrent failure modes that the Wasserstein and conditional variants were introduced to mitigate [16,31,36], and that the variational and hybrid VAE-GAN approaches address by trading a measure of sharpness for a more stable coverage of the data manifold [40–42].

Table 8. Quality-evaluation instruments used to assess the fidelity and diversity of GAN-generated outputs across three representative architectures.

Model	Quality metric used	Purpose
StackGAN	Inception Score (IS)	Measures diversity and quality via predicted class probabilities.
PGGAN	Multi-scale statistical similarity	Checks whether local image structure matches the training set across multiple scales.
Table-GAN	Information loss (L-2 norm)	Quantifies discrepancy in mean and standard deviation between synthetic and real features.

5.4. Dimension II: downstream utility for detection and augmentation

Utility concerns whether synthetic data measurably improves the downstream task, here the accuracy and robustness of intrusion and malware detectors. The reviewed evidence supports two distinct and partly opposing utility claims. The constructive claim is that class-balancing generators materially improve minority-class detection: conditional tabular

GANs and their ensemble-balanced variants raise recall on rare attack families that classifiers would otherwise ignore [16,30,33,34] and multi-module designs address high-dimensional imbalanced traffic in the same spirit [53], Wasserstein-penalised oversampling stabilises the enrichment of tabular attack data [31,36], and bidirectional and autoencoder-coupled GANs extend the same benefit to unsupervised and one-class detection regimes [37–39]. Variational and focal-loss autoencoder formulations report comparable gains on imbalanced benchmarks while offering more stable training [13,14,40,43].

The adversarial claim is that the same generative machinery can be turned against detectors. Two capabilities recovered from the corpus illustrate this dual use directly. MalGAN reliably synthesises adversarial malware examples that bypass black-box detectors [57]; crucially, those same examples can be recycled into adversarial training to harden production systems, so the capability is genuinely double-edged. IDSGAN produces adversarial variants that inflict high evasion-increase rates against traditional classifiers such as support vector machines, k-nearest-neighbours, and multilayer perceptrons [58]. The mechanism by which IDSGAN preserves attack semantics while inducing misclassification is instructive and is set out in Table 9: functional features that encode the essential behaviour of the attack are held fixed, while non-functional features are perturbed with noise, so the resulting flow remains a working attack yet is read as benign by the detector [47–49].

Table 9. IDSGAN feature-mapping strategy: functional features are preserved to retain attack semantics while non-functional features are perturbed to induce misclassification.

Feature group	Treatment and rationale
Functional features (unchanged)	Protocol type, service, flag, and attack-specific behaviours held fixed to preserve the nature of the attack.
Non-functional features (perturbed)	Intrinsic, content, time-based, and host-based features modified by random noise to induce misclassification without breaking attack functionality.

A qualification of considerable practical importance emerges when the utility of augmentation is examined from the analyst's rather than the classifier's viewpoint. Statistical inspection of GAN-generated flows indicates that a sample flagged as abnormal is not necessarily malicious: many synthetic samples that a detector marks as “not normal” lack the feature correlations characteristic of genuine attacks, for instance the coordinated patterns that define a denial-of-service event, and therefore represent statistical noise rather than a coherent adversarial threat. The implication for the field is that fidelity measured only as distributional proximity is an insufficient warrant of utility; a generator can pass an aggregate fidelity test while producing samples that are semantically empty as attacks [23,52,56]. This decoupling of statistical abnormality from operational maliciousness is one of the review's central cautions and is revisited in the Discussion.

5.5. Dimension III: security and privacy risks of synthetic data

The third dimension inverts the customary framing: rather than asking what synthetic data can do for a detector, it asks what synthetic data

reveals about the private records on which the generator was trained. The reviewed evidence establishes that generative models, and diffusion models in particular, are susceptible to membership-inference attacks that determine whether a specific record was part of the training set, a direct privacy breach when the training data are sensitive network captures or user records [47,55]. Table 10 consolidates a taxonomy of membership-inference attacks against diffusion models, reporting their access assumptions and attack success rate, as compiled from the diffusion membership-inference literature [9,21].

Table 10. Taxonomy of membership-inference attacks against diffusion models, with access assumptions and attack success rate (ASR).

Attack class	Method	Access	Applicable to DMs	ASR on DMs
Baseline	Random guess	N/A	Yes	0.500
White-box	GAN-Leaks	White-box	No	0.615 (theoretical upper bound)*
Black-box	Shadow model + LOGAN	Discriminator	Yes	0.544
Black-box	Shadow model + TVD	Discriminator	Yes	0.089
Black-box	Over-representation	Discriminator	Yes	0.532
Black-box	Monte-Carlo set	Discriminator	Yes	0.505
Black-box	GAN-Leaks	Discriminator	Yes	0.507
Query-based	LOGAN	Black-box	No	N/A
Query-based	TVD	Black-box	No	N/A
Query-based	SecMIstat	Generator	Yes	0.811
Query-based	SecMINNs	Generator	Yes	0.888

DM = diffusion model; ASR = attack success rate. *The GAN-Leaks white-box value of 0.615 is a theoretical upper bound obtained by generating precise latent codes through DDIM inversion, because standard gradient-based optimisation is computationally infeasible for diffusion models.

Two findings dominate Table 10. First, attacks designed for GANs transfer poorly to diffusion models: the classic black-box GAN-Leaks and shadow-model constructions barely exceed the random-guess baseline of 0.500, and several are not even applicable to the diffusion setting. Second, attacks that exploit the deterministic reverse–denoise structure of diffusion models, SecMIstat and its neural-network variant SecMINNs, which operate on the step-wise error introduced in Section 5.3, achieve attack success rates of 0.811 and 0.888, far above every GAN-oriented method. The privacy risk of diffusion models is therefore not inherited from earlier generative families; it is specific to the diffusion mechanism and must be measured with diffusion-aware instruments.

The severity of this leakage is quantified further in Table 11, which reports attack success

rate and area under the ROC curve for the two SecMI variants across six image datasets, as reported in the source study [9]. SecMINNs is uniformly the stronger attack, reaching an AUC of 0.956 on Tiny-ImageNet and exceeding 0.94 on every dataset examined; SecMIstat, though weaker, still attains AUCs near 0.88. The consistency of these figures across datasets of markedly different scale and content indicates that the vulnerability is a property of the diffusion training objective rather than an artefact of any single corpus.

Table 11. Attack success rate (ASR) and area under the ROC curve (AUC) for SecMI variants across image datasets; higher values indicate greater privacy leakage.

Dataset	Method	ASR (↑)	AUC (↑)
CIFAR-10	SecMIstat	0.811	0.881
CIFAR-10	SecMINNs	0.888	0.951
CIFAR-100	SecMIstat	0.798	0.868
CIFAR-100	SecMINNs	0.872	0.940
STL10-U	SecMIstat	0.809	0.881
STL10-U	SecMINNs	0.892	0.950
Tiny-IN	SecMIstat	0.821	0.894
Tiny-IN	SecMINNs	0.903	0.956
Pokemon	SecMIstat	0.821	0.891
COCO2017-val	SecMIstat	0.803	0.875

Tiny-IN = Tiny-ImageNet. SecMINNs consistently outperforms SecMIstat, confirming that a learned attack model extracts more membership signal from the step-wise error than a purely statistical test.

Aggregate accuracy metrics understate the risk to the most exposed records. Table 12 reports, from the same source study [9], true-positive rates at the stringent false-positive-rate thresholds of 1% and 0.1%, the regime that matters for high-sensitivity deployments where even a small number of confident re-identifications is unacceptable. SecMINNs identifies members with a true-positive rate approaching 38% at a 1% false-positive rate on CIFAR-10 and still 7.59% at a 0.1% false-positive rate, an order of magnitude above the statistical variant. For an intrusion-detection pipeline trained on confidential enterprise captures, a low-false-positive-rate leakage of this magnitude means that a determined adversary could confirm, with high confidence, that a particular sensitive flow was present in the training corpus.

Table 12. High-confidence membership inference: true-positive rates (TPR) at low false-positive-rate (FPR) thresholds for the two SecMI variants.

Method	Dataset	TPR @ 1% FPR	TPR @ 0.1% FPR
SecMIstat	CIFAR-10	9.11%	0.66%
SecMIstat	Tiny-IN	12.67%	0.69%
SecMINNs	CIFAR-10	37.98%	7.59%
SecMINNs	Tiny-IN	29.77%	3.50%

The corpus also reports how far standard training-time defences blunt these attacks. Table 13 evaluates two common data-augmentation techniques against SecMIstat, as reported in the source study [9]. RandomHorizontalFlip is markedly protective, lowering the attack success rate from 0.964 to 0.811 and the AUC from 0.912 to 0.881, whereas Cutout is almost inert, leaving both metrics near their undefended values. The lesson is that augmentation-based defence is real but uneven: only augmentations that genuinely perturb the reconstruction geometry the attack depends upon confer protection, and defences must therefore be validated against the specific attack rather than assumed to generalise [50,51].

Table 13. Effect of training-time data-augmentation defences on the effectiveness of the SecMIstat attack.

Defence method	ASR (SecMIstat)	AUC (SecMIstat)
No augmentation	0.964	0.912
RandomHorizontalFlip	0.811	0.881
Cutout	0.961	0.908

Finally, the review confirms that the vulnerability persists at industrial scale. Table 14 reports membership-inference metrics, drawn from the same source study [9], for two production-grade Stable Diffusion latent-diffusion models evaluated on Laion-aesthetic data. Both v1.4 and v1.5 leak measurably, with AUCs near 0.70, attack success rates near 0.66, and true-positive rates at a 1% false-positive rate above 18%. The attenuation relative to the small-scale results of Table 11 reflects the diluting effect of large, diverse training corpora, but the leakage does not vanish; privacy risk is a durable property of the diffusion paradigm rather than a small-model artefact [45,56].

Table 14. Membership-inference vulnerability of large-scale Stable Diffusion latent-diffusion models evaluated on Laion-aesthetic-5plus data.

Victim model	AUC	ASR	TPR @ 1% FPR
stable-diffusion-v1-4	0.707	0.664	18.47%
stable-diffusion-v1-5	0.701	0.661	18.58%

The mechanism that makes these diffusion-specific attacks possible is worth stating precisely, because it explains why diffusion privacy must be evaluated differently from GAN privacy. The strongest attacks exploit a deterministic reverse process Φ_θ that maps a clean sample to a latent state, paired with a deterministic denoise process Ψ_θ that maps the latent back toward the sample; both are built on the same noise-prediction estimator f_θ. The step-wise error is the squared distance between a sample and its reconstruction through this deterministic cycle, and because the reconstruction is a function of the learned weights rather than of sampling stochasticity, that error is systematically smaller for training members than for unseen samples, the very asymmetry quantified in Table 7. Deterministic reverse and denoise implementations are thus a prerequisite for reliable membership inference, and equally for principled defence [44,45].

6. Discussion

Read across its three dimensions, the assembled evidence resolves into a coherent and cautionary account of the state of generative synthetic-traffic research. Fidelity has advanced to the point where the newest generators, especially diffusion models, produce samples that are nearly indistinguishable from real data under aggregate distributional metrics [44,45,56]. Utility is real but conditional: class-balancing generators demonstrably improve minority-class detection [16,33,34,36], yet the same fidelity that makes augmentation effective also enables the synthesis of adversarial traffic capable of evading deployed detectors [47–49]. Security, the least examined of the three dimensions, emerges as the field's most urgent blind spot, because the diffusion models that best serve fidelity are precisely those most exposed to membership inference [55].

The first cross-cutting theme is an inherent tension between fidelity and privacy. The results establish that these are not independent objectives to be optimised separately but coupled quantities in fundamental opposition. The step-wise error that certifies a diffusion model as high-fidelity, with low and consistent reconstruction of the data manifold, is the identical signal that a SecMI attack exploits to confirm membership (Tables 7, 10, and 11). A generator cannot be pushed toward perfect fidelity without simultaneously sharpening the reconstruction asymmetry between members and non-members, and hence enlarging its privacy attack surface. For network-security practitioners this means that the pursuit of ever-more-realistic synthetic captures is not a free good; beyond a point it converts a data-augmentation asset into a data-disclosure liability [46,56].

The second theme is the decoupling of statistical abnormality from operational maliciousness. The finding that many GAN-generated samples flagged as “not normal” in fact lack the feature correlations of genuine attacks (Section 5.4) undermines a common but tacit assumption in the augmentation literature, that a sample which enlarges the abnormal region of feature space is thereby a useful attack exemplar. It is not. A synthetic denial-of-service flow that violates the joint distribution of packet-rate, duration, and flag features while failing to reproduce the coordinated structure of a real attack teaches a detector to fire on noise rather than on threat. Fidelity evaluation must therefore be extended from distributional proximity to semantic validity, a requirement that current metrics such as the Fréchet Inception Distance and information-loss norms do not capture [23,52,56].

The third theme concerns transferability and the maturation of the threat model. The near-baseline performance of GAN-oriented attacks against diffusion models (Table 10) shows that privacy and robustness findings do not transfer freely across generative families; each architecture demands its own evaluation apparatus. This is a methodological warning for a field that has often imported metrics and threat models wholesale from computer vision. The persistence of leakage at industrial scale (Table 14) and the uneven efficacy of standard defences

(Table 13) together indicate that the defensive literature lags the offensive one, and that claims of privacy protection cannot be accepted without attack-specific validation [50,51,55].

These themes carry direct implications for practice. Synthetic-traffic pipelines intended for release or sharing should be evaluated not only for fidelity and downstream utility but explicitly for membership-inference exposure at low false-positive-rate thresholds, since aggregate accuracy conceals the high-confidence re-identifications that matter most in security settings (Table 12). Where diffusion models are used, deterministic reverse and denoise implementations should be adopted both to audit leakage and to reason about defences, and augmentation-based mitigations should be selected for their demonstrated effect on the specific attack rather than assumed protective. Above all, the evaluation of synthetic traffic should be tri-dimensional by default: a generator that scores well on fidelity and utility but has not been assessed for security has been only two-thirds evaluated [51,56].

6.1. Limitations

Several limitations qualify these conclusions. First, because the review is protocol-driven and framework-guided rather than meta-analytic, the incommensurability of architectures, datasets, and metrics across the corpus precludes the pooling of effect sizes; the synthesis reports convergent patterns rather than a single quantitative estimate. Second, a substantial portion of the most rigorous privacy evidence is drawn from image benchmarks such as CIFAR-10 and Tiny-ImageNet rather than from native network-flow data, so the transfer of the precise leakage magnitudes to tabular intrusion-detection datasets, while mechanistically well-motivated, remains to be confirmed empirically on networking corpora. Third, the field's rapid pace means that the newest diffusion-for-tabular-data methods are represented by a small and recent set of studies whose bibliographic details warrant confirmation against the publisher record before the findings are treated as settled [44,45,56].

7. Conclusion

This article has synthesised the evidence on generative artificial intelligence for synthetic network-traffic generation in intrusion detection through the lens of a three-dimensional fidelity–utility–security framework. The synthesis shows that the field has made substantial progress on the first two dimensions: modern generators, and diffusion models in particular, achieve statistical fidelity high enough that synthetic and real samples are difficult to separate under aggregate metrics, and class-balancing generators deliver real improvements in the detection of rare attack classes [16,33,36,44]. These gains are genuine and should be consolidated.

The review's principal contribution, however, is to foreground the third and least-examined dimension. The evidence demonstrates that the very fidelity that makes synthetic traffic useful is inseparable from a privacy vulnerability specific to the diffusion mechanism: diffusion-aware membership-inference attacks such as SecMIstat and SecMINNs achieve attack success rates approaching 0.89 and remain effective at industrial scale and under several common defences, whereas the assumption of maliciousness for merely abnormal synthetic samples is shown to be unsafe [23,55,56]. Fidelity and privacy are coupled quantities in tension, and statistical abnormality is not a proxy for operational threat.

The agenda that follows is threefold. First, the community needs semantic-validity metrics that certify synthetic attacks as operationally coherent rather than merely distributionally abnormal. Second, it needs privacy evaluation adopted as a standard, tri-dimensional reporting requirement, measured with architecture-appropriate, deterministic instruments and reported at low false-positive-rate thresholds. Third, it needs a defensive literature that matches the sophistication of the offensive one, with mitigations validated against the specific attacks they claim to resist [50,51]. Delivering a reproducible protocol, a three-dimensional evaluation framework, and this security-forward

research agenda is the contribution this work sets out to make, treating fidelity, utility, and security as co-equal criteria offers the field a route to realising the benefits of synthetic traffic without importing new risks into the systems it aims to defend [56].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the Centre for Cyberspace Studies, Nasarawa State University, Keffi for institutional support. No external funding was received for this work.

References

[1] T. Ndayipfukamiye, J. Ding, D. S. Sarwatt, A. G. Philipo, H. Ning, Adversarial defense in cybersecurity: a systematic review of GANs for threat detection and mitigation, arXiv:2509.20411 (2025).

[2] R. Flood, G. Engelen, D. Aspinall, L. Desmet, Bad design smells in benchmark NIDS datasets, in: 2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P), IEEE, 2024, pp. 658–675.

[3] M. A. Talukder, M. M. Islam, M. A. Uddin, K. F. Hasan, S. Sharmin, S. A. Alyami, M. A. Moni, Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction, J. Big Data 11 (1) (2024) 33.

[4] G. Agrawal, A. Kaur, S. Myneni, A review of generative models in generating synthetic attack data for cybersecurity, Electronics 13 (2) (2024) 322.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Adv. Neural Inf. Process. Syst., Vol. 27, 2014, pp. 2672–2680.

[6] D. P. Kingma, M. Welling, Auto-encoding variational Bayes, in: Proc. 2nd Int. Conf. Learn. Represent. (ICLR), 2014.

[7] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: Adv. Neural Inf. Process. Syst., Vol. 33, 2020, pp. 6840–6851.

[8] T. Stadler, B. Oprisanu, C. Troncoso, Synthetic data – anonymisation groundhog day, in: 31st USENIX Security Symposium, USENIX Association, 2022, pp. 1451–1468.

[9] J. Duan, F. Kong, S. Wang, X. Shi, K. Xu, Are diffusion models vulnerable to membership-inference attacks?, in: Proc. 40th Int. Conf. Mach. Learn. (ICML), PMLR, 2023, pp. 8717–8730.

[10] J. Clements, Y. Yang, A. A. Sharma, H. Hu, Y. Lao, Rallying adversarial techniques against deep learning for network security, in: 2021 IEEE Symp. Ser. Comput. Intell. (SSCI), IEEE, 2021, pp. 1–8.

[11] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ 372 (2021) n71.

[12] H. Ding, Y. Sun, N. Huang, Z. Shen, X. Cui, TMG-GAN: generative adversarial networks-based imbalanced learning for network intrusion detection, IEEE Trans. Inf. Forensics Secur. 19 (2024) 1156–1167.

[13] X. Xu, J. Li, Y. Yang, F. Shen, Toward effective intrusion detection using log-cosh conditional variational autoencoder, IEEE Internet Things J. 8 (8) (2020) 6187–6196.

[14] J. He, X. Wang, Y. Song, Q. Xiang, Network intrusion detection based on conditional Wasserstein variational autoencoder with generative adversarial network and one-dimensional convolutional neural networks, Appl. Intell. 53 (2023) 12416–12436.

[15] J. Shi, M. Xu, H. Hua, H. Zhang, S. Ermon, J. Leskovec, TabDiff: a mixed-type diffusion model for tabular data generation, in: 13th Int. Conf. Learn. Represent. (ICLR), 2025.

[16] K. S. Babu, Y. N. Rao, MCGAN: modified conditional generative adversarial network for class imbalance problems in network intrusion detection system, Appl. Sci. 13 (4) (2023) 2576.

[17] Y. N. Rao, K. S. Babu, An imbalanced generative adversarial network-based approach for network intrusion detection in an imbalanced dataset, Sensors 23 (1) (2023) 550.

[18] S. Allagi, T. Pawan, W. Y. Leong, Enhanced intrusion detection using conditional-tabular-generative-adversarial-network-augmented data and a convolutional neural network: a robust approach to addressing imbalanced cybersecurity datasets, Mathematics 13 (12) (2025) 1923.

[19] N. Sivaroopan, D. Bandara, C. Madarasingha, G. Jourjon, A. P. Jayasumana, K. Thilakarathna, NetDiffus: network traffic generation by diffusion models through time-series imaging, Comput. Netw. 251 (2024) 110616.

[20] M. Meeus, F. Guépin, A.-M. Creţu, Y.-A. de Montjoye, Achilles' heels: vulnerable record identification in synthetic data publishing, in: Computer Security – ESORICS 2023, LNCS, Vol. 14345, Springer, 2024, pp. 380–399.

[21] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE Symp. Secur. Priv. (SP), IEEE, 2017, pp. 3–18.

[22] P. Kumar, A. S. M. Tayeen, S. Misra, H. Cao, J. Liu, Q. Gong, J. Harikumar, NetDiffuser: deceiving DNN-based network attack detection systems with diffusion-generated adversarial traffic, arXiv:2603.08901 (2026).

[23] S. Rahman, S. Pal, S. Mittal, T. Chawla, C. Karmakar, SYN-GAN: a robust intrusion detection system using GAN-based synthetic data for IoT security, Internet Things 26 (2024) 101212.

[24] S. Zavrak, M. Iskefiyeli, Anomaly-based intrusion detection from network flow features using variational autoencoder, IEEE Access 8 (2020) 108346–108358.

[25] H. Zhang, J. Zhang, B. Srinivasan, Z. Shen, X. Qin, C. Faloutsos, H. Rangwala, G. Karypis, Mixed-type tabular data synthesis with score-based diffusion in latent space, in: 12th Int. Conf. Learn. Represent. (ICLR), 2024.

[26] M. Arab Loodaricheh, et al., MAGE-ID: a multimodal generative framework for intrusion detection systems, arXiv:2512.03375 (2025).

[27] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, A. Anwar, TON_IoT telemetry dataset: a new generation dataset of IoT and IIoT for data-driven intrusion detection systems, IEEE Access 8 (2020) 165130–165150.

[28] D. Perez Fiadzeawu, et al., Beyond accuracy: fidelity evaluation framework for synthetic data in network intrusion detection, advance online publication (2026).

[29] D. Sun, J. Q. Chen, C. Gong, T. Wang, Z. Li, NetDPSyn: synthesizing network traces under differential privacy, in: Proc. 2024 ACM Internet Measurement Conf. (IMC), ACM, 2024, pp. 1–11.

[30] V. Kumar, D. Sinha, Synthetic attack data generation model applying generative adversarial network for intrusion detection, Comput. Secur. 125 (2023) 103054.

[31] G.-C. Lee, J.-H. Li, Z.-Y. Li, A Wasserstein generative adversarial network–gradient penalty-based model with imbalanced data enhancement for network intrusion detection, Appl. Sci. 13 (14) (2023) 8132.

[32] G. Andresini, A. Appice, L. De Rose, D. Malerba, GAN augmentation to deal with imbalance in imaging-based intrusion detection, Future Gener. Comput. Syst. 123 (2021) 108–127.

[33] M. R. A. B. Soflaei, A. Salehpour, K. Samadzamini, Enhancing network intrusion detection: a dual-ensemble approach with CTGAN-balanced data and weak classifiers,

J. Supercomput. 80 (11) (2024) 16301–16333.

[34] O. Habibi, M. Chemmakha, M. Lazaar, Imbalanced tabular data modelization using CTGAN and machine learning to improve IoT botnet attacks detection, Eng. Appl. Artif. Intell. 118 (2023) 105669.

[35] Y. Kim, Y. Kim, H. Kim, A model training method for DDoS detection using CTGAN under 5GC traffic, Comput. Syst. Sci. Eng. 47 (1) (2023) 1125–1147.

[36] J. Engelmann, S. Lessmann, Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning, Expert Syst. Appl. 174 (2021) 114582.

[37] M. Arafah, I. Phillips, A. Adnane, M. Alauthman, N. Aslam, An enhanced BiGAN architecture for network intrusion detection, Knowl.-Based Syst. 314 (2025) 113178.

[38] T. K. Boppana, P. Bagade, GAN-AE: an unsupervised intrusion detection system for MQTT networks, Eng. Appl. Artif. Intell. 119 (2023) 105805.

[39] W. Xu, J. Jang-Jaccard, T. Liu, F. Sabrina, J. Kwak, Improved bidirectional GAN-based approach for network intrusion detection using one-class classifier, Computers 11 (6) (2022) 85.

[40] Y.-D. Lin, Z.-Q. Liu, R.-H. Hwang, V.-L. Nguyen, P.-C. Lin, Y.-C. Lai, Machine learning with variational autoencoder for imbalanced datasets in intrusion detection, IEEE Access 10 (2022) 15247–15260.

[41] R. Ahmad, et al., Towards an effective intrusion detection model using focal loss variational autoencoder for Internet of Things (IoT), Sensors 22 (15) (2022) 5822.

[42] W. Tian, Y. Shen, N. Guo, J. Yuan, Y. Yang, VAE-WACGAN: an improved data augmentation method based on VAEGAN for intrusion detection, Sensors 24 (18) (2024) 6035.

[43] O. H. Abdulganiyu, T. A. Tchakoucht, Y. K. Saheed, H. A. Ahmed, XIDINTFL-VAE: XGBoost-based intrusion detection of imbalance network traffic via class-wise focal loss variational autoencoder, J. Supercomput. 81 (2025) 16.

[44] P. Wang, Y. Song, X. Wang, DIFT: a diffusion-transformer for intrusion detection of IoT with imbalanced learning, J. Netw. Syst. Manag. 33 (48) (2025) 1–24.

[45] M. Villaizán-Vallelado, M. Salvatori, C. Segura, I. Arapakis, Diffusion models for tabular data imputation and synthetic data generation, ACM Trans. Knowl. Discov. Data 19 (5) (2025) 125.

[46] S. Alabdulwahab, Y.-T. Kim, Y. Son, Privacy-preserving synthetic data generation method for IoT-sensor network IDS using CTGAN, Sensors 24 (22) (2024) 7389.

[47] E. Alhajjar, P. Maxwell, N. Bastian, Adversarial machine learning in network intrusion detection systems, Expert Syst. Appl. 186 (2021) 115782.

[48] D. Han, Z. Wang, Y. Zhong, et al., Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors, IEEE J. Sel. Areas Commun. 39 (8) (2021) 2632–2647.

[49] C. Zhang, X. Costa-Pérez, P. Patras, Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms, IEEE/ACM Trans. Netw. 30 (3) (2022) 1294–1311.

[50] K. Roshan, A. Zafar, S. B. Ul Haque, Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system, Comput. Commun. 218 (2024) 97–113.

[51] V.-T. Hoang, Y. A. Ergu, V.-L. Nguyen, R.-G. Chang, Security risks and countermeasures of adversarial attacks on AI-driven applications in 6G networks: a survey, J. Netw. Comput. Appl. 232 (2024) 104031.

[52] Y. Zhang, Q. Liu, On IoT intrusion detection based on data augmentation for enhancing learning on unbalanced samples, Future Gener. Comput. Syst. 133 (2022) 213–227.

[53] J. Cui, L. Zong, J. Xie, et al., A novel multi-module integrated intrusion detection

system for high-dimensional imbalanced data, Appl. Intell. 53 (2023) 272–288.

[54] J. Gui, Z. Sun, Y. Wen, D. Tao, J. Ye, A review on generative adversarial networks: algorithms, theory, and applications, IEEE Trans. Knowl. Data Eng. 35 (4) (2023) 3313–3332.

[55] I. Rosenberg, A. Shabtai, Y. Elovici, L. Rokach, Adversarial machine learning attacks and defense methods in the cyber security domain, ACM Comput. Surv. 54 (5) (2021) 1–36.

[56] A. X. Wang, S. S. Chukova, C. R. Simpson, B. P. Nguyen, Challenges and opportunities of generative models on tabular data, Appl. Soft Comput. 166 (2024) 112223.

[57] W. Hu, Y. Tan, Generating adversarial malware examples for black-box attacks based on GAN, arXiv:1702.05983 (2017).

[58] Z. Lin, Y. Shi, Z. Xue, IDSGAN: generative adversarial networks for attack generation against intrusion detection, in: Advances in Knowledge Discovery and Data Mining (PAKDD 2022), LNCS, Vol. 13282, Springer, 2022, pp. 79–91.