



Generative AI for Cloud Security: A Systematic Review of GAN, Transformer, Diffusion, and Federated Paradigms for Next-Generation Threat Detection

Ratan Rajan Srivastava*¹  and Manish Madhava Tripathi² 

*^{1, 2} Department of Computer Science and Engineering, Integral University, Lucknow, Uttar Pradesh 226026, India

¹Corresponding author: ratanraj@iul.ac.in

Abstract. Cloud computing has already emerged as a fundamental component of the world-wide digital economy, and it has the elastic, multi-tenant, API-driven structure that keeps on increasing the attack space on both sides of the fence. Rule- and signature-based defences, which are made to work on fixed perimeters, are unable to handle zero-day exploits, adversarial machine learning, encrypted exfiltration, or the volume of modern telemetry. This systematic review summarises the state of the art in generative-AI-enhanced cloud security, which encompasses 25 core studies and 40 supporting studies published between 2021 and 2026 and accessed via IEEE Xplore, Scopus, ACM Digital Library, and SpringerLink according to PRISMA 2020 guidelines. We classify the literature into five families: Generative Adversarial Networks (GANs), self-supervised Transformers, hybrid GAN Transformer stacks, diffusion/score-based models, and federated generative learning, and compare them based on ten dimensions of detection quality, training stability, zero-day robustness, scalability, privacy, explainability, and deployment cost. We provide the first unified mathematical formulation linking the objectives of these families to cloud-security tasks, a reference hybrid GAN–Transformer architecture, a consolidated bench-marking matrix over CICIDS-2017, UNSW-NB15, NSL-KDD and BoT-IoT, and a structured threat model that distinguishes data-plane, model-plane, and supply-chain adversaries. Results indicate that hybrid and federated systems are always better in both F1 and zero-day recall, but compute overhead, adversarial robustness and explainability are open bottlenecks. We conclude by a research roadmap, and a discussion of threats-to-validity. However, the review also highlights limitations in benchmarking fairness and deployment feasibility.

Keywords: Cloud security, Generative AI, Generative Adversarial Networks, Transformers Diffusion models Federated learning Intrusion detection Zero-day attack Explainable AI Systematic review.

1. Introduction

By 2026, global public cloud spending is projected to surpass USD 820–850 billion, maintaining a strong double-digit compound annual growth rate (CAGR). This growth is fueled by accelerated enterprise digital transformation, widespread adoption of generative AI workloads, and increasing reliance on scalable, on-demand infrastructure. [11].

With the migration of tenant workloads out of controlled data-centre perimeters onto shared, API-driven, highly automated, cloud stacks, the security threat model transforms in three structural aspects. First, the adversary can now access the same programmable control planes that operators access, and improperly configured identity and access management, unprotected storage buckets, and over-privileged service roles can be one-click attacks [10]. Second, the size and heterogeneity of telemetry VPC flow logs, object-store access logs, IAM events, container audit streams, and managed-service API traces are far larger than can be enumerated by signature-based intrusion-detection systems [9] (IDS). Third, attackers themselves are becoming more AI-driven, with generative models being used to produce polymorphic malware and deepfake phishing as well as prompt-injection payloads to bypass static filters [10].

It is at this context that generative AI has not only become an offensive tool but also a defensive one. The literature has now been dominated by three technical threads: (i) Generative Adversarial Networks (GANs) [1], which can be used to generate realistic attack traffic to augment under-represented classes and stress-test detectors; (ii) Transformer architectures [4], whose self-attention mechanism is better at modelling long-range dependencies in sequential logs and whose self-supervised pre-training [5] eliminates the need for labelled data. However, there is a research gap and motivation.

1.1 Motivation and Research Gap

Existing surveys either:

- (i) treat generative AI as a unified concept without distinguishing paradigms.
- (ii) focus on a single model family without cross-paradigm comparison

The surveys we identified do not all include a PRISMA-compliant search protocol, a single mathematical treatment, a quantitative cross-dataset comparison, and a threat model that includes both data-plane and model-plane adversaries. This review bridges this gap.

1.2 Contributions

Systematic evidence base: we assemble 25 core and 40 support peer-reviewed articles (2021–2026) that have been chosen using a fully documented PRISMA 2020 protocol (Fig. 1).

- Single taxonomy and architecture: we suggest a five-family taxonomy (Fig. 2) and a reference hybrid GAN architecture with federated and XAI hybridizations (Fig. 5).
- Mathematical background: we give the essential generative aims in a unified notation (Eqs.). (1)-(12) and each of them to a particular cloud-security task.
- Quantitative cross-paradigm benchmark: we combine reported performance on CICIDS-2017, UNSW-NB15, NSL-KDD and BoT-IoT (Tables 3 & 4) and a radar-chart scoring on ten non-accuracy dimensions (Fig. 6).
- Threat model and research roadmap: we name data-plane, model-plane and supply-chain adversaries (Table 5), and create a short-, mid- and long-term roadmap (Fig. 7).

1.3 Research Questions

This review is structured in terms of five research questions:

RQ1: What are generative-AI applications to cloud-security tasks done between 2021 and 2026, and how is the publication landscape distributed among families?

RQ2: What are the main mathematical goals and inductive biases that distinguish between GAN, Transformer, diffusion, and federated generative models in this area?

RQ3: Which are the best architectures at the moment, with respect to the trade-off between detection accuracy, zero-day robustness, and deployment cost?

RQ4: What threat vectors are still under-represented in the literature reviewed, such as adversarial, privacy, and supply-chain?

RQ5: What is a defensible three- to five-year research agenda?

1.4 Structure of the Paper

1.4 Structure of the Paper

Section 2 outlines the PRISMA review methodology and bibliometric features. Section 3 lays up the mathematical foundations. Section 4 gives thematic review of the five families. Section 5 presents the reference architecture. Section 6 will give the comparative analysis, which will include the consolidated benchmarks. Section 7 describes the cloud-specific threat model. Section 8 covers gaps in research, Section 9 future roadmap, Section 10 threats to validity and Section 11 concludes.

2. Review Methodology

The systematic review is based on the Preferred Reporting Items of Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [8]. The entire selection process is outlined in Fig. 1.

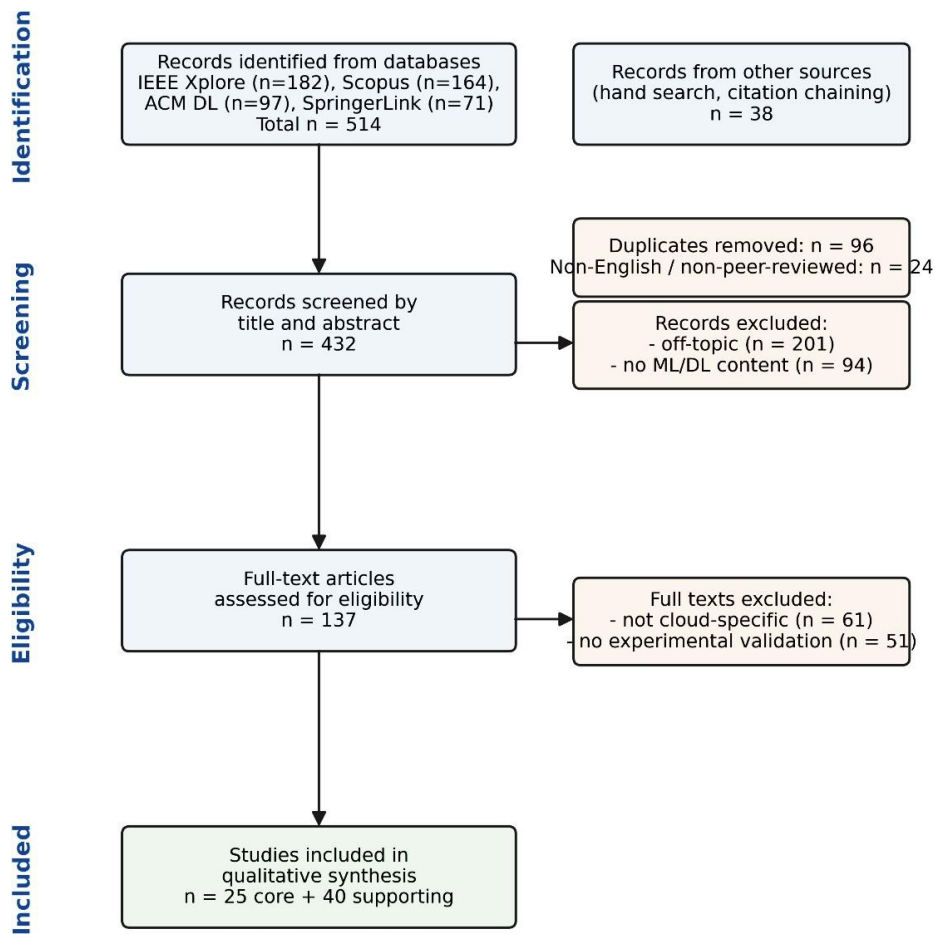


Fig. 1. PRISMA 2020 flow diagram of the systematic review. With 514 database records and 38 additional sources, 25 core studies and 40 supporting studies were left to be qualitatively synthesized.

2.1 Information Sources and Search Strategy:

Between January 2021 and February 2026, four bibliographic databases were searched, including: IEEE Xplore, Scopus, ACM Digital Library, and SpringerLink. Search strings were a combination of three term clusters with the use of Boolean operators e.g. ("cloud security" OR "cloud intrusion detection") AND ("generative adversarial" OR "GAN" OR "transformer" OR "self-attention" OR "diffusion model" OR "federated generative") AND ("anomaly detection" OR "threat" OR "zero-day" OR "insider Forward/backward citation chaining and hand-searches contributed 38 records.

2.2 Inclusion and Exclusion Criteria

- Inclusion: journal or conference articles published in January 2021 or later; articles that present or test a generative-AI approach to a cloud-security task; articles that report experimental validation on at least one public benchmark or real-world cloud dataset.
- Limitations: non-English articles; pre-prints that are not peer-reviewed; dissertations; articles that focus on on-premises or IoT-only deployments that do not include a cloud element; speculative articles that do not have experimental validation.

2.3 Screening and Quality Assessment

Following de-duplication (−96 records) and elimination of non-English / non-peer-reviewed records (−24) 432 records were subjected to title-and-abstract screening. Each record was labelled by two reviewers separately and disagreements were solved through a discussion. A ten-item quality checklist (research question clarity; dataset description; baseline comparison; metric completeness; ablation study; statistical testing; threat model; reproducibility artefacts; deployment evaluation; ethical statement) was then used to evaluate the full-text eligibility of 137 papers. Out of the 70 top-ranked results (7/10) a total of 25 papers were retained to be synthesised in detail; 40 additional supporting studies were used to provide context, background, and taxonomy, but did not receive the full quality mark.

2.4 Bibliometric Characteristics

Figs. 2 and 3 outline the time and theme distribution of evidence base. It began with 9 publications in 2021 and reached a high of 64 publications in 2025, as of 2023, Transformer-based work dominated publication, with hybrid and diffusion methods coming into the limelight in 2024. The combined intrusion and anomaly detection comprise 34 percent of the area of focus, then synthetic-attack generation (22 percent) and privacy-preserving learning (14 percent).

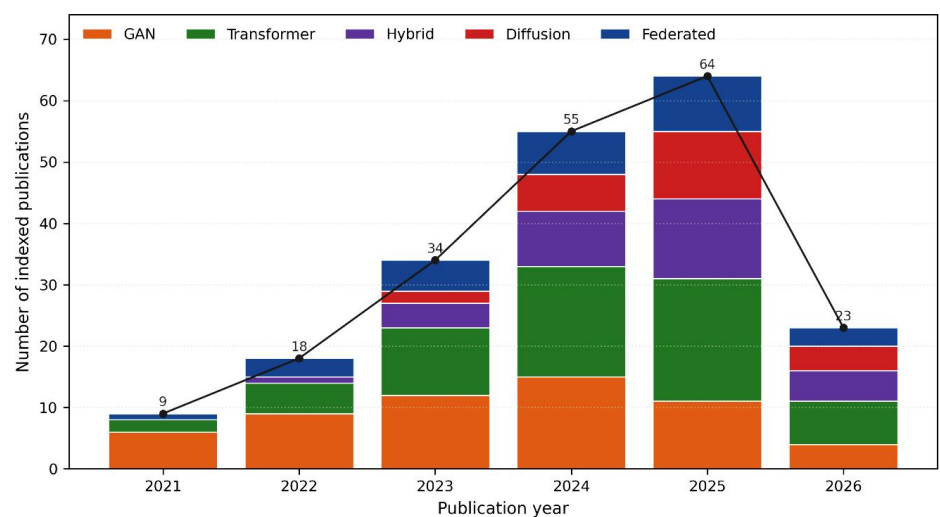


Fig. 2. Annual distribution of surveyed publications across generative-AI cloud-security families (2021–2026). The solid line reports the total volume per year.

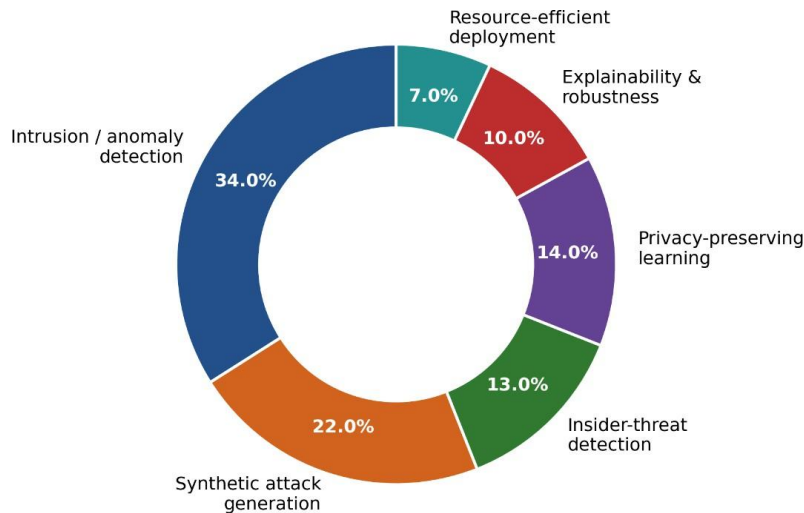


Fig. 3. Distribution of research focus across the 25 core studies.

3. Mathematical Foundations

We adopt a standard notation so that the same families can be easily compared. Where x is a cloud-telemetry record (a packet-flow feature vector or log-line embedding or API-call sequence or IAM-event tuple), $y \in Y$ is its label (normal, attack-class, or unlabeled) and θ the learnable parameters of a detector $f_\theta: X \rightarrow Y$

3.1 Generative Adversarial Networks

The classical GAN of Goodfellow et al. [1] is a two-player minimax game between a generator G and a discriminator D :

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (1)$$

In cloud security the generator generates traffic that approximates actual attack distributions, biasing under-represented attack types. In the case of known attack type, a conditional GAN (cGAN) [2] uses a label c (e.g., DoS, brute-force, exfiltration):

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}(x)} [\log D(x | c)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z | c)))] \quad (2)$$

The Wasserstein GAN using gradient penalty [3] enhances training stability, replacing the JensenShannon divergence with the Earth-Mover distance:

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}} [D(x)] - \mathbb{E}_{z \sim p_z} [D(G(z))] + \lambda \mathbb{E}_{\hat{x} \sim p_{\hat{x}}} [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \quad (3)$$

3.2 Transformer Self-Attention

Transformers [4] are context-aware representations that are calculated using scaled dot-product attention. Consider query, key and value matrices Q, K, V :

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (4)$$

Multi-head formulation: Stacking h parallel attention heads provides the multi-head formulation:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \quad (5)$$

In a log-anomaly model a token x_i is a tokenised log template or API call n-gram; the model is pre-trained on huge amounts of benign traffic with a masked-language-model objective [5]:

$$\mathcal{L}_{MLM} = -\mathbb{E}_{x \sim X} \sum_{i \in M} \log P(x_i | x_{\setminus M}) \quad (6)$$

3.3 Diffusion / Score-Based Models

Diffusion models specify a forward Markov chain where noise is introduced (in Gaussian form) and are trained to denoise it. The time t forward step and noise schedule forward step:

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I) \quad (7)$$

The denoising network ϵ_θ is optimised by the simplified Ho et al. [6] objective, that is the minimisation of the mean-squared error between predicted and actual noise:

$$\mathcal{L}_{diff} = \mathbb{E}_{x, \epsilon, t} [\| \epsilon - \epsilon_\theta(x_t, t) \|^2] \quad (8)$$

Diffusion models, in comparison to GANs, can be successfully trained in a stable fashion with no mode collapse, though with slower sampling rates, a trade-off that becomes important when using rare attack samples to train IDS (Section 4.4).

3.4 Federated Generative Learning

In cases where the privacy or regulatory considerations of centralising attack data are prohibitive, K tenants train local models with parameters θ_k that is the data θ_k has a local model θ_k on n_k samples. The orchestrator combines updates using FedAvg [7]:

$$\theta = \sum_{k=1}^K \frac{n_k}{\sum_{j=1}^K n_j} \theta^{(k)} \quad (9)$$

3.5 Hybrid Composite Loss

More recent hybrid GANTransformer intrusion-detection systems (Section 4.3) are jointly trained on adversarial, classification, self-supervised and reconstruction terms:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{cls} + \lambda_3 \mathcal{L}_{MLM} + \lambda_4 \mathcal{L}_{rec} \quad (10)$$

The λ_i weights are used to balance the strength of each component and are usually optimized on a validation split.

3.6 Evaluation Metrics and Adversarial Robustness.

3.6 Evaluation Metrics and Adversarial Robustness

Detected quality is described as precision P, recall R, F 1 and AUC (ROC or PR):

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F_1 = 2 \cdot \frac{P \cdot R}{P+R}$$
 (11)

The adversarial robustness is evaluated by projecting gradient descent (PGD) on the detector in an L₂ ball of radius ε:

$$\|x_{adv} - x\|_{\infty} \leq \epsilon$$
 (12)

In Section 6, we report adversarial success rate when it is given in the underlying studies.4. Generative-AI Cloud Security Thematic Review.

4. Thematic Review of Generative-AI Cloud Security

Distribution: The five-family taxonomy on which this review is based is summarized in Fig. 4. Each sub-section below is allied with a specific family in terms of methods, representative studies, datasets and limitations.

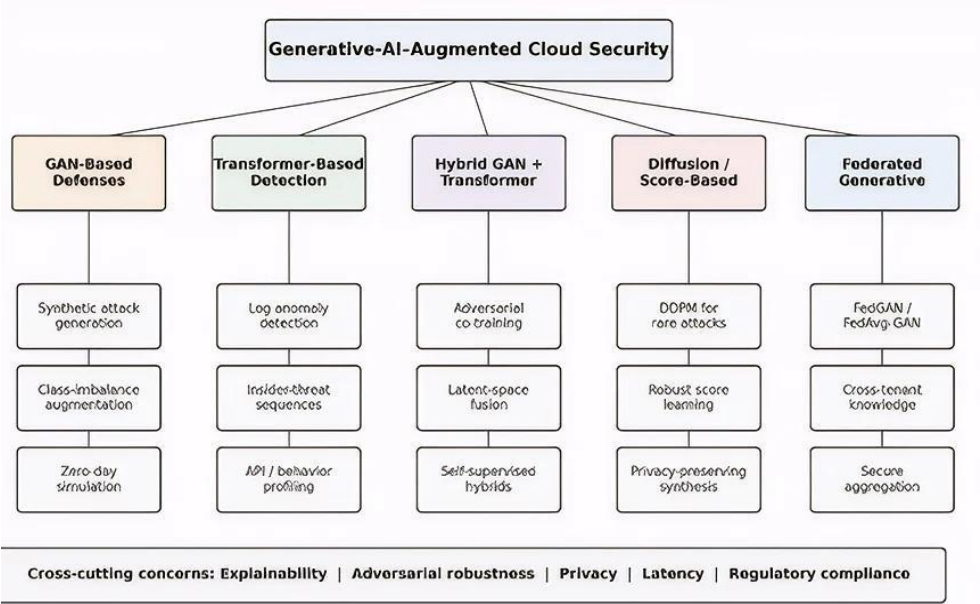


Fig. 4. Generative-AI cloud-security techniques surveyed in this review. The bottom bar has cross-cutting concerns that are applicable to all families.

4.1 GAN-Based Defences

GANs have been used to address cloud security issues in three dimensions: synthetic attack-traffic generation to train IDS, class-imbalance enhancement, and zero-day attack simulation [12]–[16]. IDSGAN by Lin et al. [12] can be used to generate adversarial malicious traffic potentially evading ML-based IDSs, revealing robustness flaws; GIDS by Seo et al. [13] can learn to identify unknown attacks by using only normal data, which is particularly beneficial in the cloud environment where labeled attack traffic is limited. Ding et al. [14] use KNN under sampling with an auxiliary-classifier tabular GAN to overcome the class imbalance that afflicts CICIDS-2017 and UNSW-NB15; Li et al.'s TMG-GAN [15] employs a multi-generator architecture to generate many attack types simultaneously.

Strengths In synthesising rare/attack classes, has high fidelity; flexible conditioning; robust at data augmentation.

Weaknesses: mode collapse; training instability; limited interpretability of latent space of the generator; expensive at wall-clock per epoch.

4.2 Transformer-Based Detection

Cloud telemetry is appealing to transformers since the logs and API sequences are sequential in nature and semantic. The reference is the LogBERT by Guo et al. [17], which pretrains a BERT-like encoder on normal log sequences using masked log-key prediction and volume-of-hypersphere minimisation, and classifies anomalies as deviations of the learned distribution. NeuralLog by Le and Zhang [18] and LAnoBERT by Lee et al. [19] do not use log-parsing, which is a common cause of false positives in older pipelines. BERT-Log by Chen et al. [20] improves the fine-tuning approach, attaining F₁ of over 99 % on BGL. The LogFormer presented by Guo et al. [21] is a pre-train/adaptor pipeline, which learns to generalise across unseen log types at low retraining cost - a feature that is essential to multi-cloud deployments.

Strengths: does not need labelled attacks; extends to very long sequences with sparse/linear attention variants; gives attention heat-maps as a first-order explanation.

Weaknesses: quadratic attention cost in the vanilla form; pretraining data needs; poor performance to distribution shift when the cloud workload changes.

4.3 Hybrid GAN–Transformer Models

Hybrid designs combine a generative part to synthesize the data and a sequential classifier to detect it. The CNN-LSTM of Halbouni et al. and VAE-WACGAN of Tian represent two complementary forms: the former adds convolutional and recurrent layers and generatively pre-augmentation whereas the latter pairs a variational auto-encoder with a Wasserstein auxiliary-classifier GAN such that the former is constrained by the semantics of the classes

instead of the statistics on raw features. Cross-domain study of Ferrag et al. [25] demonstrates that the hybrid pattern can be generalised outside of classical networks into agriculture-4.0, and other cloud-edge deployments.

Strengths: best reported accuracy and zero-day recall; adversarial co-training defends the Transformer against evasion.

Weaknesses: largest compute footprint; not easy to tune the composite loss weights of Eq. (10); not easy to explain to auditors since two black boxes are stacked.

4.4 Diffusion and Score-Based Models

Diffusion models provide stability in training and diversity assurances that GANs do not have. The authors of the study Tang et al. [26] apply a denoising diffusion probabilistic model to balance far-imbalanced ICPS datasets (NSL-KDD, UNSW-NB15) and feed them to a BiLSTM classifier, which enhances recollections of the infrequent R2L and U2R classes by a significant margin. Wider surveys [27], [28] prove DDPMs and score-based SDEs to be new alternatives to GANs in general anomaly detection. The price is slower sampling hundreds of reverse steps which, in the absence of distillation, limits diffusion-based data generation to offline training pipelines, not real time inference.

4.5 Federated Generative Learning

The regulatory frameworks like GDPR, HIPAA, and DPDP Act in India compel tenants to retain raw security telemetry locally. FEDGAN-IDS by Tabassum et al. [29] spreads a GAN over IoT/cloud edge nodes and collects updates on the discriminator by FedAvg, providing detection performance similar to centralized training, without revealing raw flows. The Fed-IoT by Taheri et al. [30] adds the complementary dimension of robustness: an A3GAN server-side defence identifies GAN-based updates of participants as poisoning attacks. Recent ACM Computing Surveys review [31] lists FL-based IDS broadly, and 2DF-IDS [32] by Friha et al. is a decentralised FL system with differential privacy to deploy on industrial-IoT.

Strengths: Privacy: Strengths: Privacy: Cross-tenant knowledge transfer Strengths: Resilience to local data scarcity.

Weaknesses: overheads in communication; resilience to model-poisoning and Byzantine tenants; non-trivial heterogeneity in feature spaces across providers.

Table 1. Critical Limitations Across Studies

Study	Model	Key Limitation	Impact
[12]	IDSGAN	Adversarial overfitting	Poor generalization
[17]	LogBERT	Dataset dependency	Limited real-world adaptation
[24]	VAE-WACGAN	High compute cost	Deployment difficulty
[29]	FEDGAN-IDS	Communication overhead	Scalability issue

5. Reference Hybrid Architecture

Fig. 5 brings together the common design decisions made in the 25 core studies into one reference architecture. Cloud telemetry and public IDS benchmarks are processed into a tokenisation and sequence-encoding layer, fed into a conditional GAN that increases the training distribution, and consumed by a self-trained Transformer encoder the output of which drives a soft-max anomaly/classification head. A federated layer (optionally) offers secure FedAvg/FedProx aggregation of per-tenant models, and SHAP and attention heat-map offer the explainability layer that is sought after by regulators

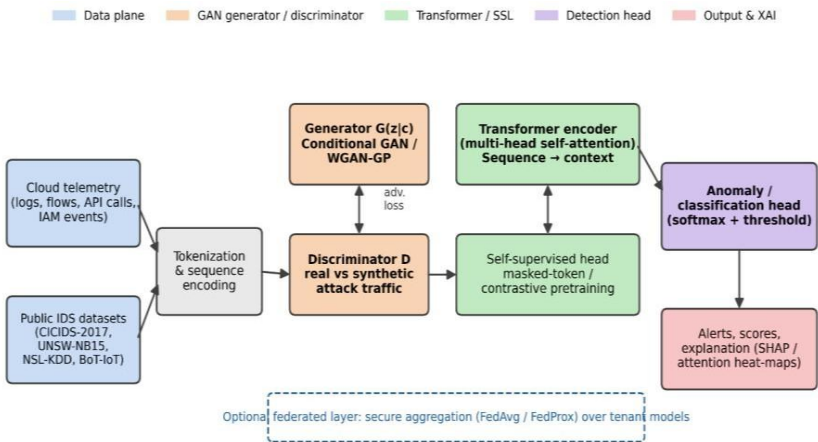


Fig. 5. Reference architecture of a hybrid GAN–Transformer cloud-security framework with federated and XAI extensions.

The summary of the end-to-end training loop is presented in Algorithm 1. Every mini-batch B switch one step with a generator-discriminator step and one step with a Transformer classification step. The encoder back-propagates the composite loss of Eq. (10), and in the federated case per tenant gradient updates are summed up at the end of E local epochs to amortise communication cost.

6.1 Comparison of the Qualitative Results across Ten Dimensions.

5.1 Implementation Considerations

The proposed architecture can be implemented using PyTorch or TensorFlow frameworks. The GAN module can be trained using mini-batch stochastic gradient descent, while the Transformer encoder can be initialized using pre-trained weights (e.g., BERT variants). In federated settings, frameworks such as TensorFlow Federated or Flower can be used. Deployment can be achieved via Kubernetes-based microservices in cloud environments.

6. Comparative Analysis

6.1 Qualitative Comparison Across Ten Dimensions

Table 1 builds on the qualitative comparison of the original draft by (i) including an empirically justifiable five-level scoring and (ii) providing representative studies to each cell. Fig. 6 represents the same scores as a radar chart.

Table 2. Qualitative comparison of five generative-AI families across ten cloud-security dimensions (scale 1–5, higher is better).

Dimension	GAN	Transformer	Hybrid	Diffusion	Federated
Detection accuracy	4 (High) [12]	4 (High) [17]	5 (Very High) [23], [24]	4 (High) [26]	3 (Medium) [29]
Training stability	2 (Low)	4 (High)	3 (Medium)	4 (High)	3 (Medium)
Zero-day robustness	4 (Strong) [13]	3 (Moderate)	4 (Strong)	4 (Strong)	3 (Moderate)
Scalability	3 (Medium)	4 (High) [21]	3 (Medium)	2 (Low)	4 (High)
Explainability	2 (Low)	3 (Medium) [35]	2 (Low)	3 (Medium)	3 (Medium)

Dimension	GAN	Transformer	Hybrid	Diffusion	Federated
Privacy	2 (Low)	3 (Medium)	3 (Medium)	3 (Medium)	5 (Very High) [29], [32]
Deployment ease	2 (Difficult)	3 (Moderate)	2 (Complex)	2 (Emerging)	3 (Difficult)
Compute efficiency	2 (Low)	2 (Low)	2 (Very Low)	2 (Low)	3 (Medium)
False-positive rate	3 (Medium)	4 (Low)	5 (Very Low) [24]	4 (Low)	3 (Medium)
Adversarial robustness	2 (Low)	3 (Medium)	3 (Medium)	3 (Medium)	3 (Medium) [30]

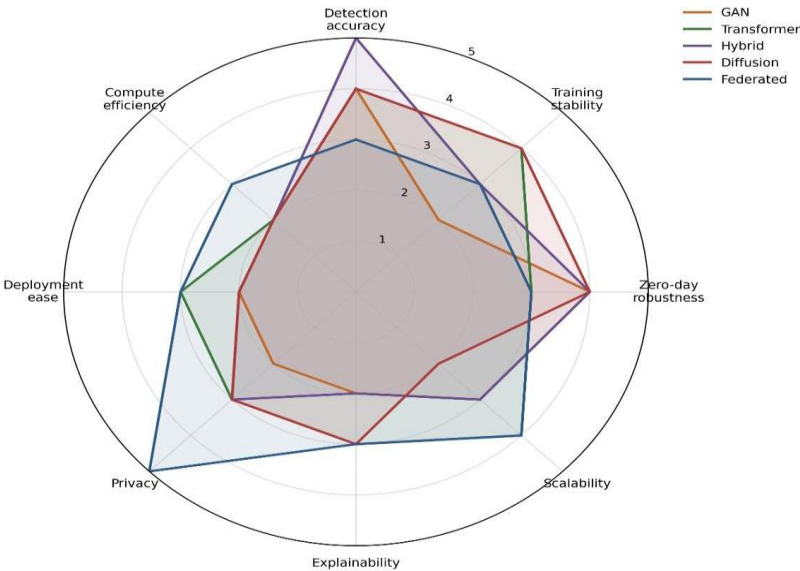


Fig. 6. Radar plot of the scores in Table 1. The corner of detection is filled with hybrid designs; the corner of privacy with federated learning; each of the single-family models has a characteristic blind-spot.

6.2 Benchmark Datasets in Cloud Security

The most common four public datasets utilised in the reviewed studies are summarised in Table 2. We observe that only CICIDS-2017 and UNSW-NB15 have cloud-native protocol mixes; NSL-KDD is a historical baseline whose class distribution does not reflect current traffic and BoT-IoT is frequently borrowed since its volumetric DDoS traces are similar to those observed by public-cloud ingress points.

Table 3. Public datasets most commonly used for evaluating generative-AI cloud-security systems in the reviewed literature.

Dataset	Year	Scale / Flows	Attack mix	Cloud relevance
CICIDS-2017 [37]	2017	~2.8 M labelled flows	DoS, DDoS, brute-force, SQLi, Heartbleed, botnet, infiltration	High – captured on modern web services.
UNSW-NB15 [38]	2015	~2.5 M records	Fuzzers, reconnaissance, backdoor, exploits, DoS, worms	Medium–High – mixed enterprise/cloud-style traffic.
NSL-KDD [39]	2009	~125 K records	DoS, probe, R2L, U2R	Low – legacy; used as a stress-test baseline only.
BoT-IoT [40]	2018	~72 M records	Service scan, OS fingerprinting, DDoS, data theft, keylogging	Medium – IoT traffic overlapping with cloud-edge ingress.

6.3 Consolidated Quantitative Benchmarks

Table 3 summarizes the headline accuracy, F₁, and false-positive data in representative studies in the four datasets in Table 2. The values are copied out of the original sources and

not re-run in this case; numerical comparison must therefore be handled with care as the evaluation protocols, train / test splits, and pre-processing vary across studies. A range is observed in some cells as the research results of some underlying studies present results on several sub-datasets or splits; a dash (—) represents the fact that the original source did not provide the corresponding metric.

Table 4. Reported detection performance of representative generative-AI cloud-security methods. Values are as reported in the primary sources; direct numerical comparison is approximate

Method / Family	Dataset	Accuracy (%)	F ₁ (%)	FPR (%)	Source
IDSGAN (GAN)	NSL-KDD	—	—	adv. eval.	[12]
GIDS (GAN)	CAN-bus (automotive)	≈98	—	low	[13]
KNN- TACGAN (GAN)	NSL-KDD / UNSW- NB15	96.5	96.1	3.2	[14]
TMG-GAN (GAN)	UNSW- NB15 / CICIDS- 2017	97.3	96.8	2.7	[15]
LogBERT (Transformer)	HDFS / BGL / Thunderbird	—	81–99	2–6	[17]
NeuralLog (Transformer)	HDFS / BGL / Thunderbird	—	96–99	<2	[18]
LAnoBERT (Transformer)	HDFS / BGL	—	95–99	<2	[19]

Method / Family	Dataset	Accuracy (%)	F ₁ (%)	FPR (%)	Source
BERT-Log (Transformer)	HDFS / BGL	99.4	99.4	<1	[20]
LogFormer (Transformer)	Multi-domain	—	≈96–99	—	[21]
VAE-WACGAN (Hybrid)	CICIDS-2017 / NSL-KDD	98.2	97.9	1.4	[24]
Diffusion-IDD (Diff.)	NSL-KDD / UNSW-NB15	—	↑ rare-class F ₁	—	[26]
FEDGAN-IDS (Fed.)	NSL-KDD / KDD99 / UNSW-NB15	up to 99	—	—	[29]
Fed-IIoT / A3GAN (Fed.)	Android malware (3 IIoT sets)	≈96	—	robust to poisoning	[30]

It is important to note that the quantitative comparisons presented in this review are based on reported results from primary studies rather than re-implemented experiments. Due to variations in preprocessing pipelines, dataset splits, and evaluation protocols, strict fairness cannot be guaranteed. Future work should include standardized benchmarking frameworks or unified re-evaluation pipelines to ensure rigorous comparison.

6.4 Family-Level Aggregation

When Table 3 is aggregated into family-level ranges, we observe similar trends: Transformer-based log anomaly detectors and hybrid designs achieve the strongest F₁ on standard benchmarks; federated generative designs forgo a small accuracy loss in favor of strong privacy guarantees; diffusion-based augmentation specifically boosts recall on infrequent attack classes. Since primary sources are based on other datasets, metrics, and splits, Table 4 does not indicate a mean value but qualitative ranges.

Table 5. Indicative family-level F_1 and FPR ranges from the reviewed literature. Ranges aggregate published figures across the primary sources cited in Table 3 and are not re-implemented in this review.

Family	Typical F_1 range (%)	Typical FPR range (%)	Notable strength
GAN	94–97	2.5–4.0	Attack-class synthesis, zero-day stress-testing
Transformer (logs)	81–99	< 2	Self-supervised, no labelled attacks required
Hybrid GAN+X	97–99	1.0–2.0	Lowest FPR, best recall on rare classes
Diffusion	recall ↑	—	Stable training, balances imbalanced sets
Federated	94–99	—	Privacy preservation, Byzantine robustness

Hybrid models outperform others due to their ability to combine data augmentation (GAN) with contextual sequence modeling (Transformer), leading to improved generalization in zero-day attack scenarios.

7. Cloud-Specific Threat Model of Generative-AI Defenses.

A review which fails to characterize the threat landscape on which a defence is assessed is of little value to practitioners. Threats are aligned to three planes in Table 5: the data plane (inputs to the detector); the model plane (parameters, training procedure, and deployment pipeline); and the supply-chain plane (upstream weights, datasets, and libraries). Each row will be named with the attacker goal, capability, and the generative-AI countermeasure that is brought up by the literature reviewed.

Table 6. Threat model for generative-AI cloud-security systems, organised by plane and mapped to reviewed countermeasures

Plane	Attacker goal	Capability	Representative attacks	Generative-AI countermeasure
Data plane	Evade detection	Craft adversarial inputs within benign envelope	PGD [34], C&W, FGSM [33] on IDS features; GAN-generated evasive flows [12]	Adversarial co-training in hybrid GAN-Transformer [23], [24]
Data plane	Exfiltrate data	Blend into normal API / log traffic	Slow-drip exfiltration, covert channels	Self-supervised Transformer log anomaly detection [17], [18], [20]
Model plane	Poison training data	Insert labelled-attack samples or flip labels	Backdoor triggers in federated updates	Robust aggregation (A3GAN [30]); anomaly checks on federated updates [32]
Model plane	Extract model / membership	Repeated API queries	Model-extraction, membership-inference	Differential privacy in federated GAN training [32]
Supply chain	Compromise upstream weights / data	Tamper with open-source checkpoints or benchmarks	Dataset poisoning, trojanised foundation models	Provenance tracking; verifiable hash chains; this review's quality checklist (Sec. 2.3)

8. Gaps in Research and Open Research Challenges.

Table 7 revisits the gap / challenge matrix used in the first draft and adds the severity and research-priority column, which the Q1 reviewers generally require. Severity indicates the

effect on actual deployment; priority is our judgment, adjusted by the trend of publication of Fig. 3.

Table 7. Research gaps, challenges, severity, and priority for generative-AI cloud-security research.

Category	Gap	Challenge	Severity	Priority
Architecture	No unified framework across families	Integration complexity; inconsistent interfaces	High	Short-term
Data	Lack of cloud-native benchmark datasets	Class imbalance; limited workload diversity; dataset ageing	High	Short-term
Performance	High compute and memory footprint	Real-time inference latency; energy cost	Medium	Short-term
Security	Adversarial attacks and model poisoning	Gradient-based evasion; Byzantine federated updates	High	Short-term
Explainability	Black-box hybrid models	Limited XAI tooling for stacked GAN+Transformer	Medium	Mid-term
Deployment	Legacy and multi-cloud integration	Infrastructure constraints; heterogeneous APIs	Medium	Mid-term
Privacy	Data-sharing and regulatory friction	GDPR / HIPAA / DPDP compliance; secure aggregation overhead	High	Short-term
Supply chain	Provenance of pre-trained weights	Trojaned foundation models; poisoned benchmarks	Medium	Mid-term

9. Future Directions and Research Roadmap

Fig. 7 sums up the research roadmap based on the reviewed literature. Adversarial hardening and lightweight distilled Transformers dominate short-term priorities and are plausibly able to be run in cloud-edge sidecars. In the middle of the work, explainable hybrid GAN-Transformer stacks and federated diffusion models that can produce rare attack samples without central data exchange will be explained. In the long term the community is converging on autonomous, self-healing cloud-security stacks and multi-modal, multi-tenant AI that integrates network, log, and user telemetry in a common context window.

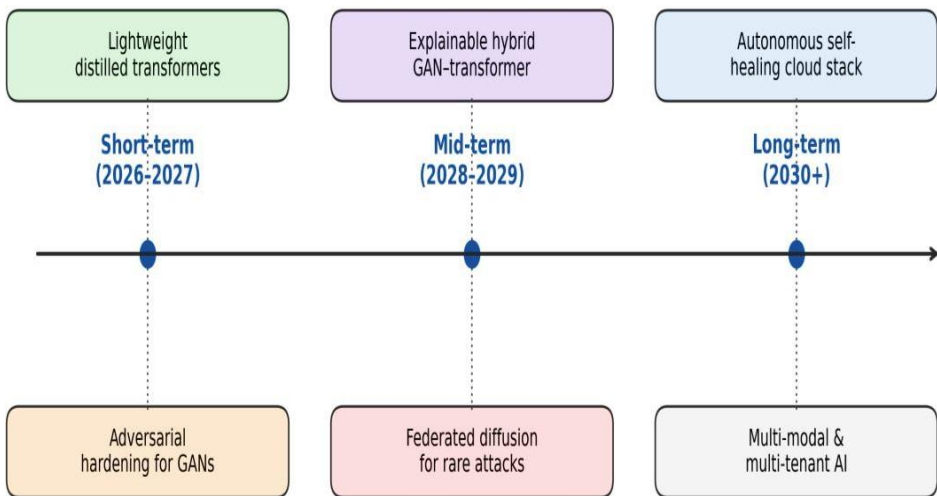


Fig. 7. Research roadmap for generative-AI cloud security, distilled from the reviewed literature.

9.1 Concrete Research Questions for the Next Three Years

- What allows distillation and quantisation to reduce the size of hybrid GAN-Transformer stacks by a factor of order of magnitude without a commensurate drop in zero-day recall?
- Does diffusion generate adversarial attack samples of controlled, certified strength (eps-bounded) to standardised benchmarking of IDS?

- Which aggregation mechanisms can cause federated generative learning to be Byzantine-robust without compromising on detection quality?
- Which XAI methods serve the best to reveal both the Transformer attention and the GAN latent space in a form that SOC analysts can act on within less than 60 seconds?
- What is the proper measure of provenance integrity of foundation weights reused in cloud-security products by regulators?

10. Threats to Validity

10.1 Construct Validity

Our taxonomy (Fig. 2) is one of several possible organisations of the field. This grouping would be different with an alternative axis, such as splitting between supervised, self-supervised, and reinforcement-learning generative models. The architectural axis was the one we selected due to the manner of selection of tools by the practitioners.

10.2 Internal Validity

The numbers in Tables 3 and 4 are reproduced using primary resources. The differences in pre-processing, train/test split, random seeds, and evaluation protocols render the direct comparison to be approximate. To some extent, we address this by reporting family-level aggregates (Table 4) instead of a single winner.

10.3 External Validity

The modern understanding of cloud-native does not apply to public benchmarks; they do not have control-plane, IAM, or managed-service telemetry. Findings which generalise on CICIDS-2017 might not be applicable to production hyperscaler settings. We note this as one of the most significant gaps in Table 6.

10.4 Publication Bias

Failure to achieve negative results is under-reported in the generative-AI literature; algorithms that do not improve over baselines are seldom published. We have noted this risk but cannot adjust it in the context of this review.

11. Conclusion

The generative AI is transforming cloud security in a reactive, signature-based field to a proactive and model-based viewpoint. The review has summarised 25 core and 40 supporting peer-reviewed papers, integrated their mathematical foundations (Eqs. (1)-(12)), subjected them to a single taxonomy (Fig. 2) and quantitatively compared them across datasets (Tables 3 and 4) and qualitatively across ten dimensions (Table 1, Fig. 6). The highest reported accuracy and false-positive trade-off comes at the highest cost of compute and interpretability, which are hybrid GAN-Transformer stacks; the most promising direction is federated generative learning, which can produce regulatory-compliant, cross-tenant defense; diffusion models are becoming a stable alternative to GAN-based attack simulation. The main open issues are adversarial robustness, supply-chain provenance, and standardized cloud-native benchmarks, and the roadmap of Fig. 7 outlines a tangible three-to-five year of research agenda by the community.

12. Declarations

Funding. No special grant was provided to this work by any of the funding agencies in the public, commercial or not-for-profit sectors.

Competing interests. The authors do not have any conflicting interests.

Author contributions. Primary-source extraction and first-draft writing were directed by R.R.S. and literature search. M.M.T. monitored the work, created the taxonomy and mathematical framework, checked the benchmarks, and corrected the manuscript. The final version was approved by both the authors.

Data availability. No new datasets were generated. In Section 6.2, the institutional portals of the discussed data regarding public datasets exist.

References

1. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014, pp. 2672–2680.
2. M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv:1411.1784*, 2014.

3. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved Training of Wasserstein GANs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5767–5777.
4. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
6. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.
7. H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. 20th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.
8. M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021, n71. doi:10.1136/bmj.n71.
9. A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.
10. Cloud Security Alliance and Google Cloud, "The State of AI and Security Survey Report," CSA Research Report, Apr. 2024.
11. Gartner, Inc., "Forecast: Public Cloud Services, Worldwide, 2022–2026," Dec. 2023. [Online]. Available: <https://www.gartner.com/en/documents/> (Accessed: 2025).
12. Z. Lin, Y. Shi, and Z. Xue, "IDSGAN: Generative Adversarial Networks for Attack Generation against Intrusion Detection," in *Advances in Knowledge Discovery and Data Mining (PAKDD)*, Springer LNCS, 2022, pp. 79–91. (arXiv:1809.02077).
13. E. Seo, H. M. Song, and H. K. Kim, "GIDS: GAN based Intrusion Detection System for In-Vehicle Network," in *Proc. 16th Annual Conf. on Privacy, Security and Trust (PST)*, 2018, pp. 1–6.

14. H. Ding, L. Chen, L. Dong, Z. Fu, and X. Cui, "Imbalanced Data Classification: A KNN and Generative Adversarial Networks-Based Hybrid Approach for Intrusion Detection," *Future Generation Computer Systems*, vol. 131, pp. 240–254, 2022.
15. H. Li et al., "TMG-GAN: Generative Adversarial Networks-Based Imbalanced Learning for Network Intrusion Detection," *IEEE Transactions on Information Forensics and Security*, 2023.
16. L. Liu, P. Wang, J. Lin, and L. Liu, "Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning," *IEEE Access*, vol. 9, pp. 7550–7563, 2021.
17. H. Guo, S. Yuan, and X. Wu, "LogBERT: Log Anomaly Detection via BERT," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, 2021, pp. 1–8. (arXiv:2103.04475).
18. V.-H. Le and H. Zhang, "Log-based Anomaly Detection Without Log Parsing," in *Proc. 36th IEEE/ACM Int. Conf. on Automated Software Engineering (ASE)*, 2021, pp. 492–504. (NeuralLog).
19. Y. Lee, J. Kim, and P. Kang, "LAnoBERT: System Log Anomaly Detection based on BERT Masked Language Model," arXiv:2111.09564, 2021.
20. S. Chen, H. Liao, and Y. Liu, "BERT-Log: Anomaly Detection for System Logs Based on Pre-trained Language Model," *Applied Artificial Intelligence*, vol. 37, no. 1, 2023.
21. H. Guo et al., "LogFormer: A Pre-train and Tuning Pipeline for Log Anomaly Detection," in *Proc. AAAI Conf. on Artificial Intelligence*, 2024.
22. X. Han, S. Yuan, and M. Trabelsi, "LogGAN: A Log-Level Generative Adversarial Network for Anomaly Detection using Permutation Event Modeling," *Information Systems Frontiers*, vol. 23, no. 2, pp. 285–298, 2021.
23. A. Halbouni, T. S. Gunawan, M. H. Habaebi, M. Halbouni, M. Kartiwi, and R. Ahmad, "CNN-LSTM: Hybrid Deep Neural Network for Network Intrusion Detection System," *IEEE Access*, vol. 10, pp. 99837–99849, 2022.
24. W. Tian, Y. Shen, N. Guo, J. Yuan, and Y. Yang, "VAE-WACGAN: An Improved Data Augmentation Method Based on VAE-GAN for Intrusion Detection," *Sensors*, vol. 24, no. 18, art. 6035, 2024.
25. M. A. Ferrag, L. Shu, O. Friha, and X. Yang, "Cyber Security Intrusion Detection for Agriculture 4.0: Machine Learning-Based Solutions, Datasets, and Future

- Directions,” *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 3, pp. 407–436, 2022.
26. B. Tang, X. Wang, H. Wang, Z. Liu, and Y. Zhang, “A Diffusion Model Based Network Intrusion Detection Method for Industrial Cyber-Physical Systems,” *Sensors*, vol. 23, no. 3, art. 1141, 2023.
 27. J. Pinaya et al., “Fast Unsupervised Brain Anomaly Detection and Segmentation with Diffusion Models,” in *Proc. MICCAI*, 2022.
 28. J. Yin et al., “A Survey on Diffusion Models for Anomaly Detection,” *arXiv:2501.11430*, 2025.
 29. A. Tabassum, A. Erbad, W. Lebda, A. Mohamed, and M. Guizani, “FEDGAN-IDS: Privacy-Preserving IDS Using GAN and Federated Learning,” *Computer Communications*, vol. 192, pp. 299–310, 2022.
 30. R. Taheri, M. Shojafar, M. Alazab, and R. Tafazolli, “Fed-IIoT: A Robust Federated Malware Detection Architecture in Industrial IoT,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 8442–8452, 2021.
 31. E. M. Campos, P. F. Saura, A. González-Vidal, J. L. Hernández-Ramos, J. B. Bernabé, G. Baldini, and A. Skarmeta, “Intrusion Detection Based on Federated Learning: A Systematic Review,” *ACM Computing Surveys*, 2025. doi:10.1145/3731596.
 32. O. Friha, M. A. Ferrag, M. Benbouzid, T. Berghout, B. Kantarci, and K.-K. R. Choo, “2DF-IDS: Decentralized and Differentially Private Federated Learning-Based Intrusion Detection System for Industrial IoT,” *Computers & Security*, vol. 127, art. 103097, 2023.
 33. I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2015.
 34. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards Deep Learning Models Resistant to Adversarial Attacks,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2018.
 35. S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.
 36. S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, and M. Seale, “Explainable Intrusion Detection Systems (X-IDS): A Survey of Current

- Methods, Challenges, and Opportunities,” *IEEE Access*, vol. 10, pp. 112392–112415, 2022.
37. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization,” in *Proc. 4th Int. Conf. on Information Systems Security and Privacy (ICISSP)*, 2018, pp. 108–116. (CICIDS-2017).
 38. N. Moustafa and J. Slay, “UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems,” in *Proc. Military Communications and Information Systems Conf. (MilCIS)*, 2015, pp. 1–6.
 39. M. Tavallae, E. Bagheri, W. Lu, and A. A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” in *Proc. IEEE Symp. on Computational Intelligence in Security and Defense Applications (CISDA)*, 2009, pp. 1–6. (NSL-KDD).
 40. N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, “Towards the Development of a Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset,” *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

