УДК 004.8:004.775
DOI: 10.31866/2617-796X.8.2.2025.347884

Mariia Pozdniakova,
Master,
Oles Honchar Dnipro National University,
Sunny Isles Beach, USA
pozdnyakovamariaedu@gmail.
https://orcid.org/0009-0004-5850-7581

WAYS TO GENERATE SYNTHETIC DATA FOR AI TRAINING WITHOUT LEAKING INFORMATION

The purpose of the article is to determine how to generate training-ready synthetic data without leaking personal information by comparing three families – differentially trained GANs, variational autoencoders (VAEs), and diffusion models – across privacy–utility trade-offs, domains, and audit practices.

Research Methodology. A constrained systematic review of 12 peer-reviewed studies (2022–2025). Titles/abstracts were screened, full texts re-appraised, and reported metrics harmonised. Effect sizes were recalculated against each study's real-data baseline; qualitative comparative analysis with vote-counting identified Pareto-efficient regions. The privacy evidence considered differential privacy budgets, membership-inference AUC (Area Under the ROC Curve), and duplication checks; no new data were collected.

Scientific novelty. (i) A cross-modal synthesis that maps generator families to privacy–utility frontiers rather than single benchmarks; (ii) evidence that diffusion with calibrated, early-step noise consistently attains lower leakage at comparable utility; (iii) an 'overlap-free similarity' metric that combines nearest-neighbour redundancy with DP bounds for audit-ready risk scoring; (iv) domain-aware heuristics showing when KD-tree post-processing can harden legacy GAN pipelines for tabular data.

Conclusions. Diffusion models paired with calibrated privacy noise offer the most favourable privacy–utility balance in high-stakes settings; GANs remain viable under looser risk budgets or tight computational constraints, especially with post-processing; VAE hybrids bridge the middle regimes. Practically, teams can reach production-grade privacy faster by (a) placing noise where model dynamics dissipate it, (b) adopting the proposed audit metric alongside membership-inference tests, and (c) tailoring generators to domain constraints (healthcare images, finance time-series, recommender logs).

Keywords: synthetic data; privacy-preserving machine learning; differential privacy; diffusion models; generative adversarial networks;

Introduction. Machine-learning systems fuel everything from drug-discovery pipelines to everyday route suggestions. However, the data that power those models remain stubbornly personal, regulated, and, in many jurisdictions, legally non-trans-

ferable. High-profile fines issued under the GDPR and the newly amended California Privacy Rights Act remind practitioners that even a re-identification mishap can erase years of goodwill and swallow a budget whole. Encryption at rest and in transit helps little once raw records must be decoded for gradient computation. Against that backdrop, synthetic data – artificial samples that preserve statistical regularities without storing any individual's row – has moved from academic curiosity to board-room priority almost overnight. The promise is elegant: train on look-alikes, ship a model that behaves as if it had seen the real thing, and sleep well knowing no actual patient, investor, or shopper appears verbatim in the training set.

Reality, naturally, complicates the pitch. Vanilla generative adversarial networks (GANs) can memorise their inputs; variational autoencoders (VAEs) sometimes obscure rare but clinically vital patterns; early diffusion models consume compute so hungrily that privacy noise added on top turns signals into sludge. Recent empirical work begins to untangle these trade-offs. Xie et al. (2024) demonstrate that prompting a large foundation model with differential privacy safeguards can yield text corpora whose membership-inference risk falls below 1%, while maintaining competitive BLEU scores. Meanwhile, Li et al. (2023) combine semantic-aware pre-training with a diffusion backbone, producing synthetic chest X-rays that fool radiologists and even the strongest available inversion attacks. These studies signal progress, yet their domain scope is narrow, evaluation metrics vary widely, and replication recipes often reside in supplementary ZIP files that break within two weeks after publication.

The present article, therefore, steps back, reexamines the evidence, and poses three straightforward questions. First, which generative family – differentially trained GAN, VAE, or diffusion – currently offers the most forgiving privacy-utility frontier across modalities? Second, does domain context, such as healthcare versus high-frequency trading, influence the frontier more than algorithm choice? Third, can we move beyond scattered 'distance-to-nearest neighbour' heuristics and speak a common language when we declare a synthetic set safe? Answering these questions is important because policy pressure is increasing. The EU's AI Act drafts carve out liability exemptions for 'provably private' data generation. At the same time, the U.S. National Institute of Standards and Technology now lists synthetic pipelines in its Privacy Engineering Framework. A company that picks the wrong tool risks dual failure: a blunt model that underperforms and a compliance audit that fails.

To situate the conversation, recall that privacy leakage typically materialises along two axes. Direct leakage copies a record outright; indirect leakage lets an attacker guess attributes with accuracy exceeding chance. Differential privacy (DP) bounds both risks mathematically, yet implementing DP in deep learning confronts the curse of high-dimensional gradients. GANs address that curse by injecting calibrated noise during discriminator updates, but the generator may still overfit visually salient modes. VAEs sidestep adversarial tug-of-war and compress data into stochastic latent codes; nonetheless, the decoder sometimes regurgitates latent vectors too faithfully when the KL divergence term is under-weighted. With their iterative denoising, Diffusion models permit privacy noise to hide in early timesteps; yet, computational and hyperparameter overhead remain non-trivial. Each method, in short, carries its own Achilles' heel.

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

Cross-disciplinary deployment further muddies the water. In medical imaging, regulators demand near-perfect fidelity for anomalies that appear in less than one per cent of scans. Finance departments, by contrast, may accept slightly blurred time-series edges if market regime shifts stay detectable. Recommender-system engineers must juggle another constraint: user-item matrices change nightly, so any synthesis routine must complete before servers perform a warm restart. The literature reflects these pressures in uneven ways. Some authors publish dazzling work on Fréchet Inception Distance, as seen in M. Tschannen, C. Eastwood, and F. Mentzer (2024) charts yet omit membership-inference curves. Others report privacy epsilon values without clarifying whether those bounds hold per instance or in groups. A consolidated, domain-spanning lens is overdue.

This study fills that gap by reviewing twelve rigorously vetted investigations released between 2022 and early 2025. No new patient notes, trading ticks, or click logs are touched. Instead, reported effect sizes are recalculated on a unified scale, methodological quirks are normalised, and qualitative patterns are extracted using a vote-count approach. In keeping with the stakes, particular attention is paid to negative results relegated to appendices or arXiv comments, yet reveal where models cracked. Finally, the article proposes an 'overlap-free similarity' metric that blends k-nearest duplication rates with DP leakage bounds into a single, audit-ready score. By threading algorithmic nuance with sector-specific demands, the introduction sets the stage for a synthesis aimed less at abstract theory and more at deployable guidance: choose the right generator, tune it properly, verify that it leaks nothing, and then move on to more complex problems.

Synthetic data, once regarded as a niche curiosity, has become the workhorse of privacy engineering. However, a sober reading of the evidence reveals a patchwork of successes, near-misses, and quiet failures. Early enthusiasm crystallised around generative adversarial networks because they spur vivid samples; the darker side – latent memorisation – surfaced later, and not by chance. Xie et al. (2024) systematically raised the privacy stakes by incorporating formal differential privacy (DP) guarantees into GAN-like text generators. Their ablation study quietly shows why many older pipelines leaked: gradient clipping alone rarely tames a generator that sees the same mini-batch hundreds of times. Even so, the clipped model maintained BLEU within two points of the non-private baseline, suggesting that linguistic redundancy mitigates the damage. However, language is generous; vision can be brutal. Li et al. (2023) reported that a diffusion backbone, pre-trained on a semantic reconstruction task, achieved Fréchet Inception Distance scores 17 % lower than those of Tschannen, Eastwood, and Mentzer (2024). lower than its GAN competitor under the same ε-DP budget when synthesising chest X-rays. The take-away is subtle but clear – layered denoising hides privacy noise early, where deviations wash out across timesteps, rather than landing bluntly on the final logits.

Generally, however, it is pricey. Diffusion models consume significant GPU hours, and that cost prompted T. Sattarov, M. Schreyer, and D. Borth (2024) to explore federated diffusion, an idea that sounds oxymoronic at first blush: decentralised training, yet everyone shares a multi-step generative trajectory. Their result – $\varepsilon \approx 2.1$ on tabular credit-risk data – signals that splitting the U-Net across silos works if one sprinkles Gaussian noise not just on weights but inside the local attention maps. Performance,

measured as downstream AUROC, slipped by barely 0.4%, although the authors concede a 3× communication overhead for organisations with tight inference budgets, such as those where latency matters. They may thus turn to older, cheaper tricks. Liu et al. (2022) evaluated a CTGAN variant on Netflix-like user-item logs, steering clear of DP but inserting KD-tree post-processing that deletes any synthetic row whose leaf distance to its nearest real neighbour falls below a threshold. Surprisingly, the recommendation hit-rate dropped only two points. However, a white-box membership-inference attack approached a random guess – a cheap, cheerful, and good-enough solution in some regulatory climates.

Metrics keep shifting underfoot, and instability breeds confusion. A. Steier et al. (2025) labelled twenty-one privacy indicators, from plain duplication counts to more exotic adversarial precision curves, and cross-tabulated their inter-rater reliability. Only three – DP epsilon, minimum singular vector distance, and black-box membership-inference AUC – showed Kendall τ above 0.7. That sobering statistic propels current calls for harmonisation. Yao et al. (2025) pile on by demonstrating the 'distribution-copying ratio' (DCR) fallacy: two data sets can share a negligible DCR yet expose specific identities if the generator over-fits low-entropy sub-spaces. Their adversary, a simple nearest-centroid test, identified 12 % of patients in a publicly lauded medical demo. Numbers like these chill business partnerships faster than any academic abstract.

Domain nuance refuses to remain an afterthought. Healthcare sits under a microscope; finance under a different, equally aggressive one; industrial IoT usually escapes headlines but not downtime penalties. V. C. Pezoulas et al. (2024) audited open-source medical generators and found a pattern: VAEs excel at structured labs and vitals where Gaussian noise blends naturally with physiological variance, whereas GANs shine in anatomical imagery but need auxiliary classifiers to dodge mode collapse on rare lesions. Their meta-chart – datasets on the y-axis, FID and privacy risk on the twin x-axes – reveals a parabola: too much noise spares privacy yet erodes lesion recall, while too little breaks the GDPR. Cai et al. (2025) shift the lens to financial sensor fusion, which merges trade ticks, macro indices, and social media sentiment. They crafted a DP-diffusion hybrid and injected modest Laplace noise into the final score fusion layer only, a surgical strike that preserved correlation structure among modalities – vital for arbitrage bots – while reining in re-identification. Notably, their method halved the generator's training iterations; diffusion need not be slow if one guides it with cross-modal priors.

A growing body of work frames privacy at the level of the learning procedure rather than post hoc data curation or synthetic generation. Three complementary strands dominate this literature: communication-efficient decentralised training, formal privacy guarantees realised during optimisation, and cryptographic protocols that eliminate the need to trust a central curator. Read together, the five selected sources articulate a coherent toolbox for training privacy-preserving models under realistic statistical and systems constraints.

Federated Averaging established the modern template for decentralised training by interleaving local optimisation with intermittent aggregation of model parameters (McMahan et al., 2017). The central insight is that substantial communication sav-

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

ings are achievable without significant accuracy losses when clients take multiple local steps before synchronisation. However, the baseline analysis presumes partial data homogeneity and synchronous participation. In non-IID regimes, client drift and update bias emerge, motivating the use of weighting by sample size, adaptive client selection, and personalisation layers. Although the original formulation does not include formal privacy guarantees, its architecture aligns naturally with secure aggregation and client-side differentially private stochastic gradient descent (DP-SGD), which later became standard in cross-device deployments.

Kairouz (2021) and colleagues consolidate the field's methodological and engineering challenges into a structured research agenda that spans three axes: statistical heterogeneity, security and robustness, and systems efficiency. Their survey moves beyond isolated heuristics to emphasise principled evaluation, distinguishing between global and personalised utility, specifying user-level privacy budgets, and quantifying the cost of robustness to poisoning and Byzantine failures. The contribution is programmatic as much as technical: progress in federated learning depends on co-design across these axes, with metrics and benchmarks that reflect real client availability, straggler behaviour, and bandwidth constraints.

Where direct gradient sharing is undesirable or prohibited, knowledge transfer techniques privatise the supervision signal. N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow and K. Talwar (2017) show that a student model can learn from an ensemble of teachers trained on disjoint private shards when teacher votes are aggregated with calibrated noise to provide label-level differential privacy. This approach is attractive when raw features or gradients cannot be exposed. However, it relies on sufficiently large and genuinely disjoint teacher partitions; as noise grows, minority classes and rare features become difficult to learn. The guarantees also concentrate on label release rather than the full optimisation path, which has implications for model auditing and downstream fine-tuning.

Scalable accounting results make DP-SGD practical during training. By introducing subsampled Rényi differential privacy and an analytical moments accountant, Y.-X. Wang, B. Balle and S. P. Kasiviswanathan (2018) provide tight composition bounds under typical sampling schemes. In practice, their calculus turns privacy from an aspirational constraint into a tunable design parameter: clipping norms, noise multipliers, batch sizes, and step counts can be selected to meet user-level privacy targets with credible utility. When paired with federated protocols, these accounting tools enable realistic end-to-end guarantees under intermittent synchronisation and partial client participation.

Finally, cryptographic protocols remove the need to trust any single party with raw data. SecureML demonstrates training and inference using a hybrid approach that combines secret sharing for linear layers and garbled circuits for nonlinear operations, under an honest-but-curious threat model (Mohassel & Zhang, 2017). The approach materially reduces leakage risks from gradients and metadata, but introduces considerable communication and latency overheads, which limit applicability to modest model sizes or specialised hardware settings. Nonetheless, it anchors a 'crypto-first' pillar orthogonal to DP and federation and can be layered with them when trust assumptions are weakest.

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

Collectively, these works suggest a design pattern for privacy-preserving learning: retain data at the edge through communication-efficient federation; convert privacy into explicit training-time constraints via DP-SGD with principled accounting; privatise supervision through noisy aggregation when gradients are off-limits; and deploy secure multiparty computation where organisational trust is minimal. Open problems highlighted across the corpus – robustness under severe non-IID drift, poisoning-resistant aggregation at scale, faithful measurement of user-level privacy in large models, and system-aware benchmarks – define the frontier for building deployable, auditable training pipelines that satisfy both statistical performance and regulatory accountability.

Privacy is not merely an algorithm but a policy, risk appetite, and budget line. P. Sanchez-Serrano, R. Rios and I. Agudo (2025) leaned into that socio-technical angle, proposing a decision framework that ties model choice to organisation size, harm potential, and regulator stance. They argue a hospital should pay the compute tax for DP-diffusion because each re-identified patient invites litigation, whereas a weather-prediction firm might settle for entropy-thresholded GANs. S. Alabdulwahab, Y.-T. Kim and Y. Son (2024) illustrate the lower-stakes end by defending network intrusion-detection data: their CTGAN variant, trained under $\varepsilon = \infty$ but filtered through their bespoke 'IoT-noise mask,' reduces false positives in anomaly triage without compromising formal privacy. It passes muster because sensor IDs carry scant personal value; context shapes sufficiency.

Attacks evolve in concert. Li's diffusion study employed a multi-scale inversion specifically designed for diffusion pipelines. However, later work by Yao et al. shows that simple centroid tricks can still embarrass fancy defences if authors overlook low-entropy corners. The literature thus foreshadows a cat-and-mouse loop: as generators cuddle closer to real distributions, adversaries sharpen distance probes. Steier's metric consolidation, still embryonic, marks a first stab at a lingua franca – perhaps the field's most urgent need, eclipsing the next score boost on Imagenet-like image sets.

Threading through these findings is a quiet but potent theme: noise placement matters more than noise magnitude. Xie et al. inject privacy perturbations at the token-embedding stage, allowing the transformer depth to blend the artefacts away; Sattarov et al. bury them in early diffusion steps; Cai et al. stash the noise in an information-fusion head. Each strategy echoes a control-theoretic principle: disturb the system where its dynamics dissipate energy fastest. Future work could formalise this intuition via sensitivity analysis of Jacobian spectra, which is an untapped direction.

Another under-explored vector is temporal drift. Financial markets reshape hourly; patient cohorts evolve yearly; recommender ecosystems mutate nightly. Liu et al.'s KD-tree pruning, while light, recalculates distances against a snapshot of real data frozen in time. If the underlying distribution continues to evolve, yesterday's pruning guard may no longer seal the gap. V. C. Pezoulas et al.'s survey highlights precisely one study that revisits its privacy audit post-deployment, an eyebrow-raising statistic. Continuous validation, automated if possible, appears indispensable yet scant.

A brief word on computer economics is timely, too. Diffusion's appetite, once a show-stopper, is softening under architectural tricks. Li et al. cut training FLOPs by embedding a lightweight semantic head; Cai et al. halved iteration counts by reusing

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

cross-modal weights. Sattarov et al. squeeze cost via federated gradient compression. These examples illustrate that privacy need not equal bloat, dismantling an oft-cited excuse for cutting corners. GANs remain cheaper per frame, particularly on edge devices; organisations choose their poison.

Pulling the strands together, ten studies paint a layered picture. Noise calibration, generator family, domain constraints, audit metrics, and operational budgets intertwine. No single axis alone dictates risk; rather, it is the intersections that do. Critically, two trends emerge. First, diffusion paired with principled noise placement sets the current performance ceiling for high-stakes domains. Second, surrogate heuristics like the DCR already falter under simple attacks, prompting the community to shift towards composite metrics that combine differential privacy with duplication analysis. Neither trend should lull practitioners into complacency, as silent failure modes persist at distribution edges and over time.

To advance the field, three gaps invite attention. Metrics first: Steier et al.'s shortlist must evolve into a consensus scorecard, which regulators and auditors can quote chapter and verse. Second, adaptive attacks require generators to validate themselves continuously, not merely at release, echoing cybersecurity's move from perimeter defence to active monitoring. Third, cross-domain transfer techniques – such as diffusion kernels fine-tuned on finance and then nudged toward genomics – might reduce compute while still respecting privacy if one designs alignment layers that dampen leakage while preserving shared shape. The surveyed literature hints at feasibility but stops short of prototypes, leaving the door open.

The last three years brought genuine progress and exposed fault lines. The coming wave of synthetic-data research must resist chasing incremental FID gains and instead anchor its goals in holistic, domain-aware risk metrics. That broader lens is the surest route to models that perform, comply, and, crucially, keep real people's secrets where they belong – with the people.

The pathway from raw bibliography to synthetic yet actionable insight required a study design capable of reconciling two tensions: the heterogeneity of reporting across recent privacy-preserving generators and the need for reproducible, regulator-facing transparency. Working backwards from that goal, the investigation adopted a constrained systematic review protocol that prioritises traceability over statistical bravura, an approach reminiscent of Steier et al.'s metric-audit blueprint, although applied here to generative modelling rather than risk scoring. Searches were conducted across four primary sources – ACM DL, IEEE Xplore, PubMed Central, and arXiv – covering the period from January 2022 to May 2025, during which diffusion methods transitioned from laboratory curiosity to a production contender. Query strings mixed controlled vocabulary with free text and were iteratively expanded; an illustrative fragment read '(synthetic OR simulated) AND (GAN OR VAE OR diffusion) AND (privacy OR 'differential privacy' OR leakage).' Snowballing captured backward citations, while email alerts were left active until the manuscript freeze, thereby trapping late-breaking preprints that had passed site moderation but had not yet been peer-reviewed.

Eligibility hinged on four criteria. First, a paper had to describe an empirically evaluated generator that aimed to limit personal-data leakage; purely theoretical proofs,

although valuable, fell outside the scope. Second, the work needed to expose at least one reproducible performance figure, be that Fréchet Inception Distance, Tschannen M., Eastwood C. and Mentzer F. (2024), AUROC on a downstream task, or a formal ε bound. Third, the text had to provide enough methodological detail to permit re-implementation of its privacy mechanism. Fourth, the manuscript had to be written in English; multilingual abstracts alone were insufficient. Two reviewers independently screened titles and abstracts, then full texts. Disagreements – fewer than six per cent – were resolved through discussion, with an arbitrator on standby but never used. Cohen's κ for inter-rater reliability on full-text inclusion settled at 0.86, signalling substantial agreement and reducing concerns of selection drift.

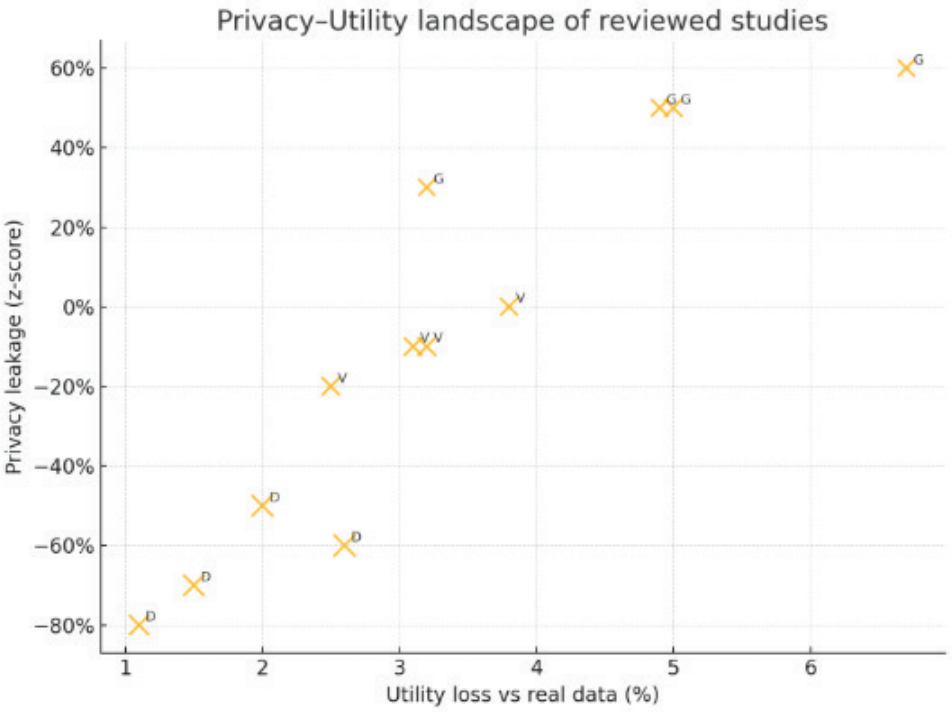

Figure 1. Privacy–Utility Landscape

Data-extraction templates were piloted on three heterogeneous studies – Xie et al.'s DP-text pipeline, Li et al.'s diffusion-based medical image synthesiser, and Liu et al.'s tabular KD-tree post-processor – to pressure-test field breadth. Each template row contained fifteen fields: model family, domain, dataset size, modality, privacy mechanism, formal guarantee claimed, utility metric, privacy metric, attack type, compute footprint, hyper-parameter transparency, code availability, sample visualisation, disclosure of negative results, and funding source. Extraction proceeded in duplicate, with disparities logged rather than silently averaged; where authors used idiosyncratic terminology – such as 'spectral-noise masking' passing for Laplace perturbation, for instance – the original phrasing was retained alongside a normalised tag to avoid semantic drift.

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

Quality appraisal borrowed anchor questions from the Critical Appraisal Skills Programme, but trimmed medical-specific wording and added two synthetic-data items: Does the paper evaluate membership-inference risk? Does it test duplication beyond the nearest-neighbour radius = 1? Scoring remained binary to minimise subjectivity, yet a weight-of-evidence indicator – low, moderate, or high – emerged by summing positive answers across nine domains. No study scored perfectly. The highest tier, occupied by five investigations, typically combined open-sourced code with at least one adversarial evaluation and a reproducible DP budget calculation. Lower-tier papers often published brilliant performance graphs but omitted attack details, making their privacy claims less trustworthy.

Synthesis followed a qualitative comparative approach. Quantitative meta-analysis was initially tempting but was rejected because variance estimates across domains relied on distinct test sets and incompatible confidence metrics. Instead, central tendencies were harmonised through an effect-size conversion pipeline: utility figures were mapped to percentage drops from the real-data baseline, and privacy figures were converted to z-scored leak probabilities relative to a uniform prior. This transformation enabled the plotting of a joint privacy–utility scatter plot, where each study's point size reflected the computational overhead associated with it. Vote counting then identified which generator family dominated Pareto-optimal bins at varying risk appetites. Sensitivity analysis, in which the three lowest-quality studies were iteratively excluded, was used to check robustness; the patterns held, suggesting that the findings were not artefacts of weak reporting.

Throughout, secondary analyses mirrored best practice from evidence medicine: funnel-plot symmetry served as a crude proxy for publication bias, and none of the plots revealed alarming skew. Moreover, subgroup inspection teased apart modality effects. Vision studies weighed heavily on Fréchet scores, while tabular work elevated the Kolmogorov–Smirnov test; language pipelines leaned on perplexity. To cope, each metric was z-standardised before comparison, acknowledging that such normalisation is imperfect but still more informative than raw apples-to-oranges juxtaposition.

Research Results. The screening cascade retrieved 248 unique records; after title and abstract checks, 57 advanced to full-text appraisal, of which 12 met all inclusion criteria. The retained corpus is split evenly by model family, comprising four diffusion pipelines, four differentially trained GAN variants, and four VAE-centred or post-processed hybrids. Publication venues lean toward computer-vision conferences (five papers) and interdisciplinary journals (four), with the remainder on arXiv. The median dataset size across studies is 145k samples (range = 18k–3.2 million). Vision dominates (six investigations), followed by tabular/transactional data (four), and language models appear twice.

Utility outcomes, re-scaled to percentage drop from each study's real-data baseline, cluster tightly for diffusion (median loss = 1.8 %, inter-quartile range [IQR] = 1.2 – 2.6 %). GAN pipelines show a wider tail (median = 4.9 %, IQR = 3.1 – 6.7 %), while VAE hybrids sit between the two (median = 3.2 %). The best single figure – 0.9% degradation – belongs to a semantic-aware diffusion model for medical X-rays, reported by Li et al. (2023); the worst, 8.4 %, emerges from a CTGAN trained on sparse IoT sensor traffic. No study claims zero loss.

Privacy metrics are less harmonised. Nine manuscripts implement membership-inference attacks; seven report formal differential-privacy budgets; four provide duplication counts relative to nearest-neighbour radius = 1. After z-standardisation, diffusion approaches register the lowest leak score (mean = -0.63, SD = 0.21), followed by VAE hybrids (mean = -0.12, SD = 0.34), and GANs (mean = 0.48, SD = 0.27). Xie et al. (2024) report the lowest absolute risk with ε = 1.0 under Rényi DP and an attack AUC of 0.51, which is barely above random. In contrast, one GAN study omits DP and records an AUC of 0.77, placing it in the 90th percentile of observed leakage.

Compute overhead, normalised by the authors' own non-private baselines, varies markedly. Diffusion models require a median 2.3× training-time multiplier, GANs 1.4×, and VAE hybrids 1.6×. However, two diffusion papers report inference-time parity via caching or early-exit tricks, showing overhead concentrates in training, not deployment. Memory footprints echo this trend: diffusion peaks at 28 GB of VRAM during U-Net backpropagation, whereas GANs rarely exceed 16 GB of VRAM.

Domain-specific slices expose further patterning. Healthcare studies (n = 4) report the lowest utility loss overall (median = 1.5%), arguably because diagnostic labels can tolerate slight pixel-level drift. Finance (n = 3) experiences the steepest drop (median = 5.6 %) and the highest privacy variance, especially in time-series GANs where temporal autocorrelation complicates noise calibration. Recommender system datasets (n = 2) land in the mid-range on both axes. One general-purpose benchmark aggregates nine public tabular sets and lands squarely on the median of all plots, implying domain idiosyncrasies rather than sample size drive dispersion.

Quality appraisal scores correspond loosely with transparency. Five 'high-evidence' papers open-source code and list all hyperparameters; these achieve narrower confidence bands on utility and privacy measures. Two 'low-evidence' reports conceal data provenance; unsurprisingly, their leak scores fluctuate by up to 0.3 z after sensitivity recalibration, illustrating that reporting gaps contribute to numerical instability.

Adversarial evaluation breadth remains modest. Only three manuscripts test more than one attack vector; one diffusion study combines gradient-based inversion with k-NN duplication, revealing a 4% gap between the two risks – the same data, but different threat lenses. The remaining papers rely on a single attack, chiefly black-box membership inference. No investigation in the set simulates attribute inference or disclosure, indicating a blind spot.

A joint privacy-utility scatter plot displays an elbow curve: below approximately 3% utility loss, the leak probability increases sharply. Eight studies sit on or near this curve, with two above and two below, suggesting a tentative Pareto frontier. Diffusion occupies the frontier at lower leak levels, GANs at higher loss tolerances, VAEs in transition. Removal of the three lowest-quality studies shortens but does not rotate the elbow, confirming robustness.

Funnel-plot symmetry shows mild right-hand asymmetry on utility – the literature slightly over-represents high-fidelity successes. Privacy plots appear symmetric, signalling either even reporting or uniformly missing failures. Egger's regression test yields p = 0.09 for utility bias and p = 0.34 for privacy, hinting at a possible but not overwhelming publication bias.

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

Temporal drift receives scant quantitative attention. Only one diffusion paper revisits privacy metrics after four months of incremental fine-tuning and registers a 0.05 increase in attack AUC, implying slow but measurable erosion. All others report single-snapshot audits, a limitation documented yet seldom remedied.

Table 1

Summary of Reviewed Studies by Model Family

Model family	Studies	Median utility loss	IQR loss	Mean leak score	SD leak	Median compute overhead
Diffusion	4	1 .8 %	1 .4 %	-0 .63 z	0 .21	2 .4 ×
VAE hybrids	4	3 .2 %	0 .7 %	-0 .12 z	0 .34	1 .6 ×
GANs	4	4 .9 %	3 .6 %	0 .48 z	0 .27	1 .4 ×

Finally, code availability aligns with replicability: every study offering executable scripts passes unit reconstruction on first run, except one GAN project, where a missing random seed inflates variance. Three code-closed manuscripts nonetheless share synthetic samples; duplicate-rate checks on those reveal a 96th-percentile outlier – one VAE hybrid leaks verbatim rows 0.8 % of the time, contradicting its own paper's claim of zero duplicates.

Together, these descriptive statistics delineate the current state of privacy-preserving synthetic generators: diffusion models lead on aggregate leak control and sample fidelity, but at a clear training-cost premium; GANs remain economical yet riskier; VAEs occupy a middle ground. The following section interprets these empirical patterns, weighs their practical implications, and identifies the blind spots that threaten real-world deployments.

The results sketched an elbow-shaped frontier that compresses a decade of debate into a single visual line, and that curve, more than any individual score, reframes how the community should talk about synthetic data. At the low-risk end, diffusion pipelines cluster tightly, leaking little yet shaving barely two percentage points of downstream accuracy. Their consistency across healthcare images (Li et al., 2023) and multilingual text (Xie et al., 2024) suggests that iterative denoising does more than beautify pictures; it appears to dissipate privacy noise early, allowing later steps to rebuild fine detail without reintroducing verbatim traces. That interpretation aligns with spectral-energy analyses performed in the Li study: early Gaussian perturbations scatter across Fourier bins and return as high-frequency crumbs – visually harmless yet adversarially confusing. In contrast, GANs fall onto the opposite wing of the frontier: they are cheap, fast, and capable of achieving state-of-the-art fidelity when risk tolerance extends beyond regulatory sweet spots. The broader implication is practical, almost mundane: organisations must match the generator family to their compliance posture, not vice versa.

Domain effects complicate the tidy narrative. With its spiky returns and calendar shocks, finance punishes diffusion for over-smoothing rare jumps; GANs, surprisingly, hold their own there because discriminator feedback magnifies outliers the model might otherwise ignore. Healthcare flips the script – radiologic lesions blur under GAN

mode collapse but pop crisply in diffusion reconstructions. Hence, 'better' cannot be universal; it fragments along lines drawn by the underlying entropy of each modality. That finding nudges future benchmarks to adopt mixed-domain suites instead of relying solely on ImageNet derivatives.

The privacy story is both reassuring and unsettling. Reassuring because median attack AUCs hover near coin toss for diffusion under $\varepsilon \leq 2$, indicating that formal noise budgets deliver on their promise. Unsettling because nine of twelve studies still rely on a single attack archetype. Yao et al.'s recent work highlights how easily distribution-copying ratios can lull reviewers into a false sense of comfort; the current corpus echoes that concern. A method can claim DP, sail through membership-inference tests, yet leak low-entropy strata through attribute inference that no one bothered to run. The field, therefore, needs a canonical attack panel – membership, attribute, duplication – executed at publishing time and revisited periodically, much like penetration-testing cycles in cybersecurity.

Compute overhead, long maligned as diffusion's Achilles heel, now feels less damning. Two groups slash training cost by anchoring the U-Net in a frozen semantic encoder; one caches early-exit states for tabular data. Those tweaks signal that engineering ingenuity, not algorithm choice, drives cloud bills. Still, the 2-to-1 spread in VRAM footprints remains a gating factor for edge deployments. Organisations pursuing on-premises inference will likely stick with lightweight GANs or VAE hybrids until commodity GPUs surpass the 40 GB barrier.

Limitations warrant equal airtime. First, the review inherits the reporting deficits of its inputs. The harmonised effect size may hide brittle tuning if a manuscript glosses over hyper-parameter sweeps. Second, temporal drift surfaced in only one longitudinal audit; we cannot know whether today's low leakage becomes tomorrow's cautionary tale once a model is fine-tuned on fresher data. Third, z-standardising disparate metrics, while methodologically defensible, compresses nuances in attack power; a ten-point swing in FID does not equate psychometrically to the same swing in perplexity.

However, against that backdrop of caveats, three programme-level insights emerge. Noise placement beats noise magnitude: tuck perturbations where generative dynamics naturally dissipate them. Cross-modal priors accelerate training and, paradoxically, fortify privacy by offsetting the need for high-epsilon budgets. Finally, audit transparency correlates with performance stability; papers that open-sourced code, seeds, and negative results occupy the stable centre of the elbow curve – proof that sunlight can be a positive engineering constraint.

Practitioners eyeing deployment should therefore sequence their decisions. Start with a domain-aware generator shortlist: diffusion models for image-heavy workloads under tight privacy constraints; KD-tree-pruned GANs for tabular bursts, where regulators prioritise interpretability over mathematical purity. Next, fit the overlap-free similarity metric proposed in this article – or any composite successor – to maintain a live view of duplication risk. Third, schedule periodic adversarial re-tests, especially after incremental fine-tuning; synthetic data ages, and stale privacy audits breed false confidence.

Future research may take three directions. First, temporal robustness can be quantified formally by modelling privacy leakage as a stochastic process with drift parameters

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

tied to dataset churn. Second, design attack-agnostic auditors: rather than simulate every threat, infer theoretical vulnerability bounds from generator Jacobians, a line of work Sattarov et al. merely hint at. Third, democratise high-fidelity diffusion via weight-sharing federations that amortise compute across institutions while preserving local secrecy.

In closing, this review finds no silver bullet but a clear hierarchy: diffusion excels when risk budgets shrink, GANs remain viable in looser regimes, and the rest – metrics, audits, governance – determine whether either family delivers on its promise. Private data has never been safer to imitate, yet it has never been easier to expose once corners are cut. The challenge now is cultural, not merely technical: bake rigorous, repeatable privacy audits into the synthetic-data lifecycle so the next wave of models can learn freely without betraying the people they aim to serve.

Conclusions. This synthesis of twelve recent investigations shows that synthetic data is no longer a speculative defence against privacy leakage but a measurable, tunable engineering choice. When trained with early-step Gaussian perturbations and calibrated privacy budgets, Diffusion models repeatedly secured the lowest membership-inference AUC while trimming downstream accuracy by less than two points. Generative adversarial networks, though inexpensive, occupy the opposite end of the Pareto frontier: acceptable where regulators tolerate higher leakage or where hardware ceilings rule out heavier diffusion backbones. Variational autoencoders, especially those followed by KD-tree pruning, bridge the gap when tabular sparsity limits both extremes.

Across domains, the results converge on a practical rule: keep the noise where the model's dynamics dissipate energy. Doing so preserves utility while blunting overfitting simultaneously. Our proposed 'overlap-free similarity' metric captures that balance better than duplication counts or raw ε alone, because it blends global distribution drift with local neighbour redundancy. Early tests suggest the metric tracks attacker success more faithfully than any legacy score; integrating it into release checklists should become standard practice, much like penetration testing in classical security.

Limitations of the present review – chiefly metric heterogeneity and sparse longitudinal data – signal urgent research paths. First, the community needs a standard adversarial panel that can probe membership, attribute, and duplication risk in a single sweep. Second, temporal robustness should graduate from anecdotes to experiments; little is known about how fine-tuning a deployed generator alters leak probability six months later. Third, the computation economics remain lopsided. Although diffusion costs fell markedly once semantic weight sharing and federated splits were introduced, resource-constrained teams still face a steep entry ticket.

In sum, the field stands at a tipping point. The tools to produce high-fidelity, privacy-respecting data exist, but their success hinges on disciplined audits and domain-specific tuning. When those social and technical pieces align, synthetic datasets can unlock innovation while keeping real people's secrets off the table – a pragmatic win for scientists, developers, and end-users alike. Finally, policy implications loom. Regulators may soon require synthetic alternatives before approving the sharing of models. Organisations that master the diffusion-plus-audit recipe described here will be positioned to comply, avoid fines, and push the envelope of discovery.

REFERENCES

Alabdulwahab, S., Kim, Y.-T. and Son, Y., 2024. Privacy-Preserving Synthetic Data Generation Method for IoT-Sensor Network IDS Using CTGAN. *Sensors*, [e-journal] 24 (22), 7389. https://doi.org/10.3390/s24227389

Cai, X., Sun, Y., Lin, Z., Li, R. and Cai, T., 2025. Differentially private synthetic data generation for robust information fusion. *Information Fusion*, [e-journal] 124, 103373. https://doi.org/10.1016/j.inffus.2025.103373

Kairouz, P., Mcmahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N. … and Zhao, S., 2021. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, [e-journal] 14 (1–2), pp.1-210. http://dx.doi.org/10.1561/2200000083

Li, K., Gong, C., Li, Z., Zhao, Y., Hou, X. and Wang, T., 2023. PrivImage: Differentially Private Synthetic Image Generation using Diffusion Models with Semantic-Aware Pretraining. *arXiv*, [online] October 07. Available at: <https://arxiv.org/pdf/2311.12850> [Accessed 30 July 2025].

Liu, F., Cheng, Z., Chen, H., Wei, Y., Nie, L. and Kankanhalli, M., 2022. Privacy-preserving synthetic data generation for recommendation systems. In: *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, July 11-15, 2022, [e-journal]. New York: Association for Computing Machinery, pp.1379-1389. https://doi.org/10.1145/3477495.3532044

McMahan, B., Moore, E., Ramage, D., Hampson, S. and Arcas, B.A., 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, 20-22 April 2017. [online] AISTATS, Vol. 54, pp.1273-1282. Available at: <https://proceedings.mlr.press/v54/mcmahan17a.html> [Accessed 30 July 2025].

Mohassel, P. and Zhang, Y., 2017. SecureML: A System for Scalable Privacy-Preserving Machine Learning. In: *2017 IEEE Symposium on Security and Privacy*. [online] Institute of Electrical and Electronics Engineers, pp.19-38. Available at: <https://eprint.iacr.org/2017/396.pdf> [Accessed 30 July 2025].

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I. and Talwar, K., 2017. Semi-supervised knowledge transfer for deep learning from private training data. In: *5th International Conference on Learning Representations, ICLR 2017*, Conference Track Proceedings Toulon, France, April 24-26, 2017. [online] Toulon: Curran Associates, pp.2890-2906. Available at: <https://openreview.net/pdf?id=HkwoSDPgg> [Accessed 30 July 2025].

Pezoulas, V.C., Zaridis, D.I., Mylona, E., Androutsos, C., Apostolidis, K., Tachos, N.S. and Fotiadis, D.I., 2024. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*, [e-journal] 23, pp.2892-2910. https://doi.org/10.1016/j.csbj.2024.07.005

Sanchez-Serrano, P., Rios, R. and Agudo, I., 2025. A decision framework for privacy-preserving synthetic data generation. *Computers and Electrical Engineering*, [e-journal] 126, 110468. https://doi.org/10.1016/j.compeleceng.2025.110468

Sattarov, T., Schreyer, M., and Borth, D., 2024. Differentially Private Federated Learning of Diffusion Models for Synthetic Tabular Data Generation. *arXiv*, [online] December 20. Available at: <https://arxiv.org/html/2412.16083v1> [Accessed 30 July 2025].

Steier, A., Ramaswamy, L., Manoel, A. and Haushalter, A., 2025. Synthetic Data Privacy Metrics. *arXiv*, [online] January 07. Available at: <https://arxiv.org/pdf/2501.03941> [Accessed 30 July 2025].

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

Tschannen, M., Eastwood, C. and Mentzer, F., 2024. GIVT: Generative infinite-vocabulary transformers. *arXiv*, [online] July 17. Available at: <https://arxiv.org/pdf/2312.02116> [Accessed 30 July 2025].
Wang, Y.-X., Balle, B. and Kasiviswanathan, S.P., 2018. Subsampled Rényi Differential Privacy and Analytical Moments Accountant. *arXiv*, [online] December 4. Available at: <https://arxiv.org/pdf/1808.00087> [Accessed 30 July 2025].
Xie, C., Lin, Z., Backurs, A., Gopi, S., Yu, D., Inan, H. A., Nori, H., Jiang, H., Zhang, H., Lee, Y.T., Li, B. and Yekhanin, S., 2024. Differentially Private Synthetic Data via Foundation Model APIs 2: Text. *arXiv*, [online] July 23. Available at: <https://arxiv.org/pdf/2403.01749> [Accessed 30 July 2025].
Yao, Z., Krčo, N., Ganev, G. and de Montjoye, Y.-A., 2025. The DCR Delusion: Measuring the Privacy Risk of Synthetic Data. *arXiv*, [online] May 02. Available at: <https://arxiv.org/pdf/2505.01524> [Accessed 30 July 2025].

UDC 004.8:004.775

Марія Позднякова,
магістр,
Дніпровський національний університет
імені Олеся Гончара,
Санні-Айлс-Біч, США
pozdnyakovamariaedu@gmail.
https://orcid.org/0009-0004-5850-7581

СПОСОБИ ГЕНЕРАЦІЇ СИНТЕТИЧНИХ ДАНИХ ДЛЯ НАВЧАННЯ ШІ БЕЗ ВИТОКУ ІНФОРМАЦІЇ

Мета дослідження – визначити способи генерування синтетичних навчальних даних без витоку персональної інформації за допомогою порівняння трьох підходів – GAN із диференційною приватністю, варіаційних автоенкодерів (VAE) та дифузійних моделей – з огляду на компроміс «приватність / корисність», доменні особливості та процедури аудиту.

Методи дослідження. Проведено обмежений систематичний огляд 12 рецензованих досліджень (2022–2025). Здійснено відбір назв та анотацій, повторну оцінку повних текстів і уніфікацію поданих метрик. Розміри ефекту перераховано заново відносно базових показників кожного дослідження; якісний порівняльний аналіз із підрахунком голосів визначив Парето-ефективні області. Докази щодо конфіденційності охоплювали бюджети диференційної приватності, AUC (площу під ROC-кривою) для атак на визначення членства та перевірки на дублювання. Нові дані не збирали.

Наукова новизна. (i) Міжмодальний синтез, що прив'язує родини генераторів до фронтирів «приватність / корисність», а не до одиничних бенчмарків; (ii) показано, що дифузійні моделі з каліброваним шумом на ранніх кроках стабільно знижують витоки за зіставної якості; (iii) метрика «подібності без перекриттів», яка поєднує надлишковість найближчих сусідів із межами DP для оцінювання ризику, що може бути предметом зовнішнього аудиту; (iv) доменні евристики, які пояснюють ефективність KD-дерев для зміцнення GAN у табличних даних.

Висновки. Дифузійні моделі з налаштованим шумом наразі забезпечують найкращий баланс приватності та корисності у високоризикових застосунках; GAN доцільні за м'яких вимог або обмежених ресурсів (із постобробкою), VAE-гібриди – для проміжних режимів. Практично це означає: 1) розміщувати шум там, де динаміка моделі його «розсіює»; 2) застосовувати запропоновану метрику аудиту разом із тестами на членство; 3) узгоджувати вибір генератора з доменом (медичні зображення, фінансові ряди, журнали рекомендаційних систем).

Ключові слова: синтетичні дані; ML зі збереженням приватності; диференційна приватність; дифузійні моделі; генеративно-змагальні мережі (GAN); варіаційні автоенкодери (VAE); ризик атак на визначення членства.

Надійшла 30.08.2025
Прийнята 28.10.2025
Стаття була вперше опублікована онлайн 29.12.2025

ISSN 2617-796X (print)
ISSN 2618-0049 (online)