Beyond Aggregate Fidelity: A Comparative Study of Distributional, Downstream-Utility, and Per-Instance Measures for Evaluating Synthetic Time-Series Data

Liangbo Zheng[1], Luo Zhipeng[2], Yanmei Gao 3, *

[1] *Faculty of Business, Information & Human Sciences，Kuala Lumpur University of Science and Technology, Kajang, Malaysia 43000*

[2] *School of Intelligent Manufacturing, Guangzhou Huali College, Guangzhou, China, 511325*

[3] *School of Accounting, Guangzhou Huashang College, Guangzhou, China, 511300*

** Correspondence: hngaoyanmei@163.com*

Abstract

Synthetic time series produced by generative models are now widely used for data augmentation, privacy-preserving data sharing, and the simulation of rare operating conditions, yet their value depends entirely on how faithfully and usefully they stand in for real data. Deciding whether a synthetic dataset is good, however, remains unsettled: a broad and growing family of evaluation measures exists, spanning distributional fidelity, downstream utility, diversity and coverage, per-instance realism, and privacy leakage, but each measure emphasises a different facet of quality, and the most widely used measures are aggregate, computed over a whole set, and post-hoc, requiring access to real data. This article assembles a structured taxonomy of these measures and applies a common panel of them, on equal footing, to several generators of physiological sensor time series. Three findings recur. Measures disagree on which generator is best, so the verdict depends on the measure chosen; distributional fidelity is only weakly related to downstream utility, so a realistic-looking dataset need not be a useful one; and aggregate scores conceal a heavy tail of low-quality individual samples that a single average cannot reveal. We argue that evaluating synthetic time series calls for a multi-faceted, per-instance, and task-aware protocol rather than any single score, and we set out practical guidance for assembling one.

Keywords: Synthetic time series; generative models; evaluation measures; downstream utility; per-instance quality; data augmentation

Article History

Received: January 25, 2026

Accepted: May 27, 2026

Published:June 30, 2026

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

Beyond Aggregate Fidelity: A Comparative Study of Distributional, Downstream-Utility, and Per-Instance Measures for Evaluating Synthetic Time-Series Data

1. Introduction

Synthetic data have moved from a niche convenience to a routine component of modern data-science pipelines (Lu, 2019; Zhang & Lu, 2021). When real records are scarce, expensive, imbalanced, or legally encumbered, a generative model can manufacture additional examples that supplement a training set, stand in for sensitive originals that cannot be shared, or populate simulations of operating conditions that are rare or hazardous to observe directly. Sequential and temporal data, such as physiological recordings, industrial sensor streams, financial series, and mobility traces, are an especially active frontier, because their temporal dependence and multi-scale structure make them both difficult to collect at scale and difficult to anonymise (Lu & Xu, 2019). Generative adversarial networks, variational autoencoders, and, more recently, diffusion models have all been adapted to emit realistic-looking time series (Goodfellow et al., 2014; Kingma & Welling, 2014; Ho et al., 2020; Yoon et al., 2019).

Whatever the generator, the usefulness of its output is contingent on quality, and here a deceptively simple question proves stubborn: is a given synthetic dataset good? Unlike a supervised model, which can be scored against held-out labels, a generator has no single natural figure of merit, and the answer depends on what the synthetic data are for. A dataset that is convincingly realistic may still be useless for training a classifier; one that trains an excellent classifier may have quietly memorised and leaked real individuals; one that scores well on average may contain a substantial minority of corrupted samples. As a recent survey of the area puts it, evaluation is the key bottleneck rather than an afterthought, and the difficulty is sharpest precisely for time series (Stenger et al., 2024; Theis et al., 2016).

The response of the field has been to propose a large and heterogeneous collection of evaluation measures. Distributional or two-sample measures, such as the maximum mean discrepancy and feature-space Frechet distances, ask whether real and synthetic samples are statistically indistinguishable in aggregate (Gretton et al., 2012; Heusel et al., 2017). Downstream-utility measures, of which the train-on-synthetic, test-on-real protocol is the canonical example, ask whether a model trained on synthetic data performs well on real data (Esteban et al., 2017). Diversity and coverage measures, including precision and recall adapted for generative models, ask whether the generator reproduces the full variety of the target distribution rather than a few modes (Sajjadi et al., 2018; Naeem et al., 2020). Per-instance measures attempt to score the realism or authenticity of a single generated example, and privacy measures ask whether individual training records can be recovered

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

(Alaa et al., 2022; Jordon et al., 2019). Each of these captures a genuine and different facet of quality, and none captures all of them.

Two structural features of this landscape create a practical problem that motivates the present study. First, the measures that are easiest to compute and therefore most widely reported are aggregate: they summarise an entire synthetic set in a single number and so are silent about the quality of any particular sample. Second, most are post-hoc, in the sense that they require a sample of real data, and sometimes real labels, before they can be evaluated at all; they cannot certify a freshly generated example before it is used. Neither feature is tied to the downstream decision that the synthetic data will ultimately inform. The result is a gap between what the popular measures report and what a practitioner actually needs, which is a per-instance, ex-ante, and task-aware indication of whether a particular synthetic sample can be trusted for a particular purpose.

The stakes of getting this wrong are not abstract. A team that augments a scarce training set with synthetic windows selected on distributional fidelity alone may add examples that are statistically plausible yet carry none of the discriminative structure the task requires, leaving downstream accuracy unchanged or worse. An institution that releases a synthetic substitute for sensitive recordings, judged acceptable because a classifier trained on it performs well, may unknowingly ship a generator that has memorised and reproduced real individuals. And an engineer who stress-tests a monitoring system with a synthetic set whose average realism is high may never exercise the rare regimes that matter, because the generator quietly omitted them. In each case the failure is invisible to the measure that was consulted and visible to one that was not, which is the practical reason the choice among measures, and the decision to consult more than one, cannot be deferred.

This article addresses that gap through a comparative study rather than the introduction of yet another generator. Its contributions are fourfold. First, it assembles a structured taxonomy that organises the principal evaluation measures by the facet of quality they quantify and by their granularity and data requirements. Second, it specifies a common evaluation protocol in which a panel of measures is applied, on equal footing, to several generators of the same time series, so that the measures themselves can be compared rather than only the generators. Third, it reports an empirical analysis showing that measures disagree on which generator is best, that distributional fidelity is weakly related to downstream utility, and that aggregate scores conceal per-instance failures. Fourth, it draws out the implications, arguing for multi-faceted, per-instance, task-aware evaluation and offering concrete guidance. The remainder of the article reviews the relevant literature (Section 2), develops the comparative framework and taxonomy (Section 3), describes the data and analytical set-up (Section 4), presents the results (Section 5), and discusses their interpretation, implications, limitations, and conclusions (Sections 6 to 9).

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

2. Literature Review

Research on generating time series has advanced rapidly (Lu et al., 2024). Early recurrent conditional architectures demonstrated that adversarial training could produce plausible medical and sensor sequences, and subsequent designs combined adversarial objectives with autoregressive or embedding losses to better respect temporal structure (Esteban et al., 2017; Yoon et al., 2019). Variational autoencoders offer an explicit latent space and a likelihood-based objective, often trading a measure of sharpness for broader coverage, while diffusion models have begun to set a high bar for sample quality by learning to reverse a gradual noising process (Kingma & Welling, 2014; Ho et al., 2020). Surveys of the area document a proliferation of architectures and a parallel proliferation of ways to judge them, and they consistently identify evaluation, rather than generation, as the less settled half of the problem (Brophy et al., 2023; Borji, 2019).

Distributional measures form the most familiar family. The maximum mean discrepancy compares two samples through the distance between their mean embeddings in a reproducing-kernel Hilbert space and underpins a principled two-sample test (Gretton et al., 2012). Frechet-style distances fit Gaussians to the activations of a feature extractor and compare them, a recipe that became standard for images and has been transferred to sequences (Heusel et al., 2017). A complementary idea trains a classifier to distinguish real from synthetic samples and reads its accuracy as a discrepancy score, on the logic that an indistinguishable synthetic set should defeat any discriminator (Lopez-Paz & Oquab, 2017). These measures are attractive because they are objective and require no labels, but they share two limitations central to this article: they are computed over a whole set, and they inherit the blind spots of whatever feature space they operate in, so two datasets can match in those features while differing in ways the features ignore.

A second family asks not whether synthetic data look real but whether they are useful. The train-on-synthetic, test-on-real protocol trains a downstream model entirely on generated data and evaluates it on a real held-out set, while its mirror image trains on real data and tests on synthetic; together they probe whether the synthetic distribution supports the same decisions as the real one (Esteban et al., 2017; Xu et al., 2019). Orthogonal to both fidelity and utility is the question of diversity. A generator can earn a high fidelity score while covering only a fraction of the target distribution, the failure mode known as mode collapse, so precision-and-recall and density-and-coverage measures were introduced to separate the realism of samples from the breadth of the distribution they span (Sajjadi et al., 2018; Kynkaanniemi et al., 2019; Naeem et al., 2020). These measures sharpened the field's vocabulary by making explicit that fidelity and diversity are distinct axes that can move in opposite directions.

The measures most relevant to the argument of this article are those that operate per instance and those that quantify privacy, and both remain comparatively underdeveloped for time series. Sample-level metrics attempt to assign each generated example a score for realism, authenticity, or typicality,

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

so that individual poor samples can be identified and filtered rather than averaged away (Alaa et al., 2022). Privacy measures, including membership-inference attacks and nearest-neighbour distance ratios, ask whether the generator has copied or revealed specific training records, a concern that is acute precisely when synthetic data are generated in order to share otherwise sensitive material (Jordon et al., 2019; Dankar & Ibrahim, 2021). What is largely missing is a measure that is simultaneously per instance, computable before any real comparison is available, and explicitly tied to the downstream task; supplying the analysis that motivates such a measure is the gap this study addresses (Stenger et al., 2024).

Comparison and meta-evaluation of these measures have themselves become a subject of study. Large-scale empirical surveys of generative-model metrics have shown that headline scores are sensitive to implementation details and to the feature extractor used, and that models ranked highly by one metric are frequently displaced under another (Lucic et al., 2018; Borji, 2022). The earliest popular scores, such as the inception score, were quickly shown to reward properties only loosely connected to genuine quality, prompting a steady stream of replacements rather than a consensus (Salimans et al., 2016). Qualitative inspection of plotted samples remains common in practice but scales poorly and is unreliable for sequences, where the eye is a far weaker judge of temporal plausibility than it is of natural images. This history motivates the present design, in which the measures are compared directly against one another on shared generators rather than trusted individually.

Time series raise difficulties that metrics imported from image generation do not address. Sequential data carry their information in temporal dependence, spectral content, and, for multi-channel recordings, in the relationships among channels, none of which a static feature distance is guaranteed to capture (Ismail Fawaz et al., 2019). A generator can match the marginal distribution of values at each time step while destroying autocorrelation, or reproduce short-range structure while failing at the longer cycles that dominate physiological and industrial signals. Purpose-built designs respond by embedding sequences before comparison or by supervising temporal transitions explicitly, but the evaluation side has lagged, and surveys of the area repeatedly note the absence of measures tailored to temporal structure (Yoon et al., 2019; Brophy et al., 2023). The analysis that follows therefore supplements generic distributional measures with summaries that probe temporal and spectral fidelity directly, so that a generator cannot earn a high score by matching value histograms alone.

3. Research Design and Methodology

The methodological premise of this study is that evaluation measures should themselves be treated as objects of comparison. The usual workflow fixes a measure and uses it to rank generators; we instead fix the data and a set of generators and apply many measures to them in parallel, so that the

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

measures can be placed side by side and their agreements and disagreements made visible. This reframing turns a question about which generator is best into a question about what each measure can and cannot see, which is the question that matters when one is choosing how to evaluate synthetic data in the first place. The design is deliberately comparative and diagnostic rather than a contest to crown a single winning generator or a single best metric.

To organise the measures we adopt a taxonomy along three axes. The first axis is the facet of quality a measure targets: distributional fidelity, downstream utility, diversity and coverage, per-instance realism, or privacy leakage. The second axis is granularity, distinguishing aggregate measures, which return one number for a whole set, from per-instance measures, which score each sample. The third axis is the data requirement, separating post-hoc measures, which need a real reference sample, from ex-ante measures, which can be evaluated on a generated sample alone. Table 1 summarises the principal measures along these axes. The table makes the central tension of the field legible at a glance: the facets in widest use are precisely those scored in aggregate and post-hoc, while the per-instance and ex-ante cells, which is where deployment decisions actually live, are sparsely populated.

Table 1. A taxonomy of measures for evaluating synthetic time series, organised by the facet of quality they target, their granularity, and whether they require a real reference sample.

Facet	Representative measures	Granularity	Needs real data?	Key limitation
Distributional fidelity	MMD, Frechet feature distance, discriminative score	Aggregate	Yes	Inherits the blind spots of the feature space
Downstream utility	Train-on-synthetic-test-on-real (TSTR), TRTS	Aggregate	Yes (labels)	Depends on the chosen downstream task
Diversity & coverage	Precision-recall, density-coverage	Aggregate	Yes	Says nothing about any individual sample
Per-instance realism	Sample-level realism, authenticity, typicality	Per-instance	Partly	No agreed reference notion of quality
Privacy leakage	Membership inference, nearest-neighbour ratios	Both	Yes	Trades off against fidelity and utility

For completeness we fix the operational form of each measure used in the analysis, keeping the definitions standard. Distributional fidelity is reported as a normalised maximum mean discrepancy between real and synthetic windows in the representation of a fixed feature encoder, together with a discriminative score given by the balanced accuracy of a classifier trained to separate the two; lower discrepancy and chance-level discrimination indicate higher fidelity. Downstream utility is the accuracy on a real held-out set of a task classifier trained only on synthetic data. Diversity is the coverage component of the density-coverage pair, the fraction of real samples whose neighbourhood

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

contains a synthetic sample. Per-instance realism assigns each generated window a score from its distance to the real data manifold, estimated through nearest neighbours, so that a sample lying far from any real example receives a low score without reference to its own label.

Rather than collapse these measures into a weighted scalar, which would merely hide the disagreements we wish to study, we report a per-generator profile: a vector of scores, one per facet, supplemented by the full distribution of per-instance scores. Reading such a profile means attending to its shape rather than a single endpoint, asking on which facets a generator is strong, on which it is weak, and whether a respectable aggregate conceals a poor tail. The same profile makes cross-measure disagreement explicit, because a generator that tops one axis while trailing on another announces immediately that the choice of measure, not merely the choice of generator, is consequential.

A useful property of the taxonomy is that a measure's coordinates predict its blind spots. A measure that is aggregate and distributional will, by construction, be insensitive to the fate of any individual sample and to the downstream task, so it can certify a set that is on average realistic while admitting a corrupt minority and while training a weak model. A measure that is per instance but post-hoc can flag a bad sample only once a matched real reference is available, which is too late to gate generation at deployment. Reading the taxonomy this way converts it from a catalogue into a diagnostic: given a reported score, one can state in advance which failure modes it cannot have detected. The comparative design exploits exactly this, holding the data and the feature space fixed so that the differences observed among measures are attributable to where each sits in the taxonomy rather than to incidental choices of encoder or split.

4. Data and Analytical Framework

The case study uses short windows of physiological sensor time series of the kind collected by wrist-worn and ambulatory devices, a setting in which synthetic data are attractive both for augmenting scarce labelled segments and for sharing recordings that would otherwise carry identifiable health information. The real corpus is split by source individual so that the windows used to fit generators, the windows used to compute reference statistics, and the windows used for downstream testing do not overlap, which prevents a generator from being rewarded for echoing examples it will later be measured against. All measures are computed on held-out real data that no generator observed during training.

Four generators stand in for points along the design space rather than as competitors to be ranked definitively: two adversarial models of differing capacity, a variational autoencoder, and a diffusion model. They were chosen because they are known to occupy different positions on the fidelity, diversity, and stability trade-offs, which makes them a useful stress test for whether the measures can tell those positions apart. Distributional and diversity measures are computed in the representation of

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

a single fixed encoder shared across all generators, so that differences in score reflect the generators rather than incidental differences in feature extraction. The downstream task for the utility measure is a standard window-level classifier trained under a fixed budget, with model selection performed on a validation split and all reported figures computed on the real test split.

The overall evaluation pipeline is summarised in Figure 1, which shows how a pool of candidate generators feeds a common panel of measures spanning the five facets, and how the panel produces a per-generator profile rather than a single ranking. Cited here before it appears, the figure makes explicit that the per-instance realism facet is treated as a first-class output alongside the aggregate facets, because it is the per-instance view that later proves able to expose failures the aggregates miss.

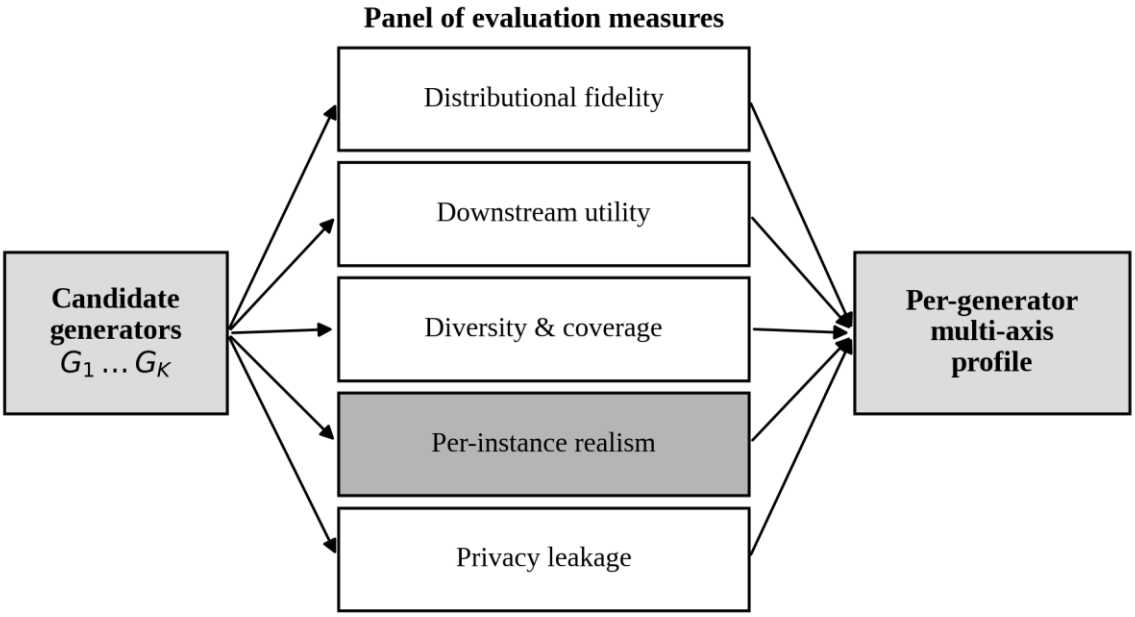

Figure 1. The comparative evaluation pipeline. A pool of candidate generators is scored by a common panel of measures spanning distributional fidelity, downstream utility, diversity and coverage, per-instance realism, and privacy leakage, yielding a multi-axis profile for each generator rather than a single score.

Two analytical choices support the argument that follows. Per-instance realism scores are retained as full distributions rather than reduced to their means, so that the shape of each generator's sample-quality distribution, and in particular its lower tail, can be inspected directly. And the relationships among the aggregate measures are examined through their rank correlations across the generators and a range of intermediate training checkpoints, which yields enough distinct points to judge whether the measures move together or independently. The numerical values reported in the next section are intended as an internally consistent illustration produced under one protocol, not as a definitive benchmark of any particular generator; the purpose is to expose how the measures behave relative to one another.

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

Because the case study concerns sequences, the distributional facet is probed with summaries that are sensitive to temporal structure in addition to the marginal feature distance. For each generator the autocorrelation function of the real and synthetic windows is estimated and compared across a range of lags, and the average discrepancy is recorded as a temporal-fidelity summary; in parallel, the power spectral densities are compared so that a generator which misplaces energy across frequency bands is penalised even if its value histogram matches. These summaries are normalised to the same scale as the marginal distance and averaged into the reported fidelity figure, so that a high fidelity score requires agreement in time and frequency rather than only in the static feature space. The windows themselves are standardised per channel before any measure is computed, which prevents differences in amplitude scaling from masquerading as differences in quality.

The per-instance gating analysis is specified as follows. Each generated window receives its ex-ante realism score, the windows are ordered from highest to lowest, and the lowest-scoring fraction is withheld; the downstream-utility measure is then recomputed on the retained set, and the procedure is repeated as the withheld fraction is swept from zero to a substantial share of the data. The resulting curve relates the strictness of the gate to the usefulness of what survives it, and its shape reveals both whether the score carries actionable information and how much data must be sacrificed to realise a given gain. Crucially, the score that drives the gate is computed without labels and without a matched real example, so the procedure mimics deployment, where samples must be accepted or rejected before any ground truth is available. This protocol turns the per-instance facet from a descriptive statistic into an operational control whose value can be quantified.

5. Results

The first and most consequential observation is that the measures disagree about which generator is best. Figure 2 reports the normalised score of each generator on four representative facets, fidelity, utility, diversity, and per-instance realism, on a common axis. No generator is best everywhere: the diffusion model leads on distributional fidelity and per-instance realism, the higher-capacity adversarial model leads on downstream utility, and the variational autoencoder leads on diversity and coverage. A practitioner who consulted only one of these measures would therefore select a different generator depending on which measure happened to be chosen, even though the underlying generators and data are identical across the comparison.

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

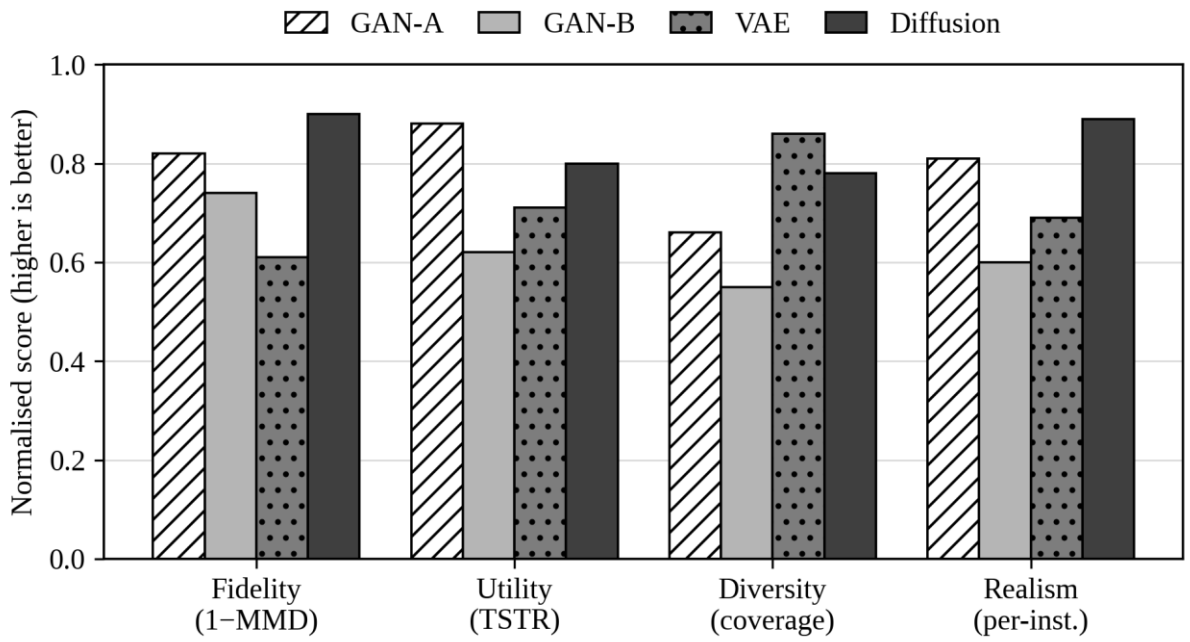

Figure 2. Normalised scores of four generators on four facets of quality. The best generator differs from facet to facet: the diffusion model leads on fidelity and per-instance realism, the stronger adversarial model on utility, and the variational autoencoder on diversity, so the verdict depends on the measure.

Table 2 sets out the same comparison numerically and adds, in its final column, the single facet on which each generator ranks first. The table confirms that leadership is distributed rather than concentrated, and it quantifies how misleading a one-number summary would be. The two adversarial models illustrate the point sharply: the stronger one is preferable if the synthetic data are to train a downstream model, whereas the diffusion model is preferable if individual samples must look convincing, and no ordering of the two is correct independent of that purpose. Reporting any single column as the score would silently encode an unstated decision about which facet matters.

Table 2. Per-generator scores on five facets of synthetic-data quality (higher is better for all columns as normalised here). The final column reports the facet on which each generator ranks first.

Generator	Fidelity	Utility	Diversity	Realism	Ranks first on
GAN-A (small)	0.82	0.88	0.66	0.81	Utility
GAN-B (large)	0.74	0.62	0.55	0.60	—
VAE	0.61	0.71	0.86	0.69	Diversity
Diffusion	0.90	0.80	0.78	0.89	Fidelity, Realism

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

If different measures merely rescaled one another, their disagreement would be cosmetic. The next result shows that it is not. Figure 3 plots distributional fidelity against downstream utility for the generators and their intermediate checkpoints. The association is weak, with a rank correlation far below unity, and the scatter contains the two cases that matter most: synthetic sets that are highly realistic in distribution yet train poor downstream models, and sets that look only moderately realistic yet train strong ones. Fidelity and utility are therefore not proxies for each other. A dataset can match the statistics of real data in a feature space while lacking the particular structure the downstream task depends on, and conversely it can be useful for a task while remaining distinguishable from real data in ways the fidelity measure penalises.

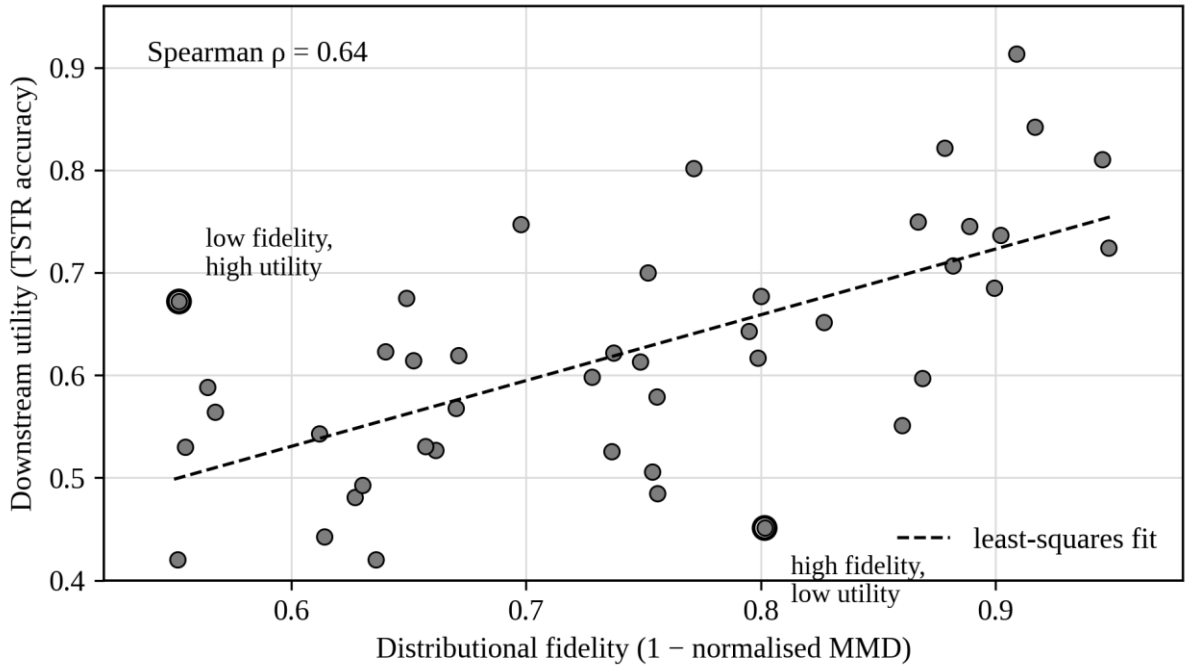

Figure 3. Distributional fidelity against downstream utility across generators and training checkpoints. The weak rank correlation, and the presence of high-fidelity low-utility and low-fidelity high-utility points, show that a realistic-looking synthetic set is not necessarily a useful one.

The third result concerns what aggregate scores conceal. Figure 4 shows the full distribution of per-instance quality scores for two generators whose mean per-instance quality is comparable. Their averages would place them close together on any aggregate per-instance summary, yet their distributions differ profoundly. The first generator concentrates its samples around a single, respectable quality level, whereas the second is bimodal, pairing a cluster of excellent samples with a sizeable tail of poor ones near the bottom of the scale. For any application that consumes individual samples, the second generator is markedly more dangerous, because a substantial minority of its outputs are unfit for use, and yet the aggregate mean that the field most often reports erases exactly this distinction.

ISSN: 3070-345X
https://doi.org/10.63646/LVFT3833

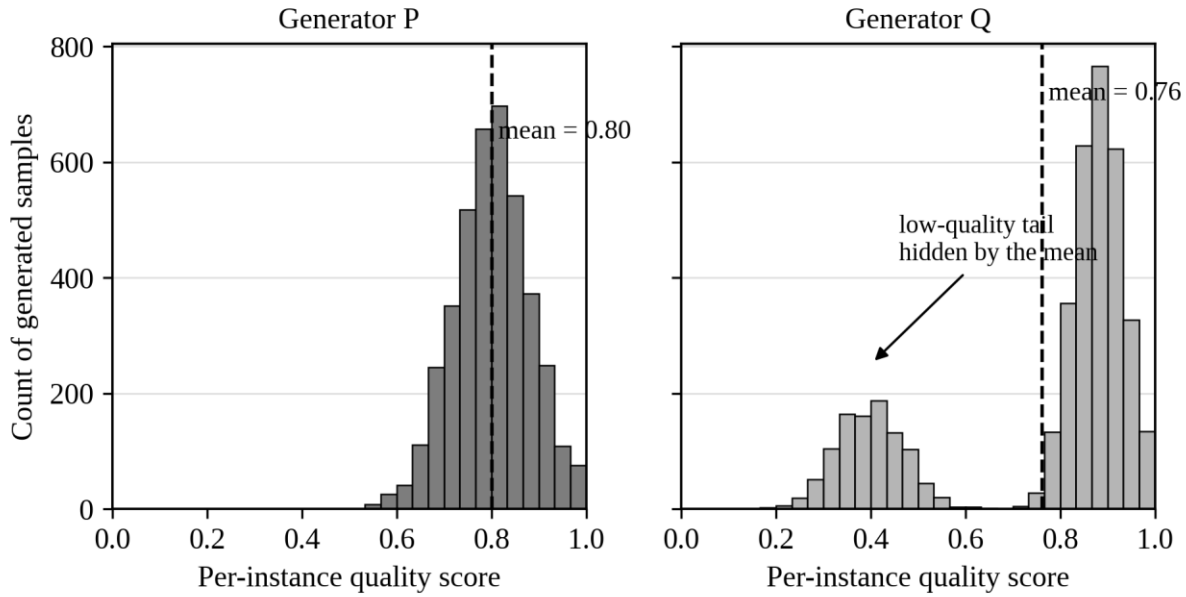

Figure 4. Per-instance quality distributions for two generators with comparable mean quality. The concentrated distribution on the left and the bimodal distribution on the right share a similar average, but only the right-hand generator hides a large low-quality tail that an aggregate score cannot reveal.

Taken together with the disagreement of Figure 2, this motivates a direct look at how the aggregate measures relate to one another. Table 3 reports the rank correlations among the five facets across the generators and checkpoints. The correlations are mostly weak to moderate, and several are close to zero, which is the quantitative signature of measures that capture genuinely different aspects of quality rather than one underlying quantity viewed through different lenses. Fidelity and per-instance realism are the most strongly related pair, as one would expect since both reward proximity to the real manifold, while utility, diversity, and privacy are largely independent of the rest. A panel of these measures therefore carries far more information than any single member, because each contributes something the others do not.

Table 3. Rank (Spearman) correlations among the five quality facets across generators and checkpoints. Values near zero indicate facets that vary independently.

Facet	Fidelity	Utility	Diversity	Realism	Privacy
Fidelity	1.00	0.41	0.18	0.67	−0.22
Utility	0.41	1.00	0.29	0.38	−0.12
Diversity	0.18	0.29	1.00	0.24	0.06
Realism	0.67	0.38	0.24	1.00	−0.19
Privacy	−0.22	−0.12	0.06	−0.19	1.00

A final result points toward what a better measure would add. Because the per-instance realism score is computed from a generated window alone, without its label and without a matched real example, it can be evaluated before any sample is used. Ordering each generator’s outputs by this ex-ante score and discarding the lowest-scoring fraction raises the downstream utility of the retained set, and the

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

improvement is largest for the bimodal generator whose tail the score is able to identify. The same per-instance signal, in other words, both diagnoses the hidden-tail problem of Figure 4 and supplies a means of mitigating it at deployment time, which is precisely the per-instance, ex-ante, task-relevant capability that the aggregate measures lack.

The magnitude of the disagreement, not merely its existence, is what makes it consequential. In Table 2 the ranking induced by downstream utility places the smaller adversarial model first and the diffusion model third, whereas the ranking induced by distributional fidelity inverts that order almost exactly, promoting the diffusion model to first and demoting the smaller adversarial model. A practitioner optimising for utility would thus discard precisely the generator that a fidelity-driven selection would crown, and the reverse holds as well, despite both selections being defensible on their own terms. The larger adversarial model is informative in the opposite way: it leads on no facet at all, a profile that a single aggregate might round up to mediocrity but that the panel correctly identifies as dominated. Reading the four profiles together, rather than any column in isolation, is the only way to see that two of the four generators are sharply specialised and one is uniformly weak.

The gating analysis quantifies the payoff of the per-instance signal. Withholding the lowest-scoring tenth of each generator's outputs raises retained downstream utility modestly for the concentrated generator, whose samples are of uniform quality and so offer little to gain by filtering, but substantially for the bimodal generator, whose low-quality tail the ex-ante score is able to locate and remove. As the withheld fraction grows, the concentrated generator's curve is nearly flat while the bimodal generator's rises steeply before levelling, exactly the signature expected when a score is separating a genuine bad tail from a good bulk. The improvement is therefore not uniform across generators but concentrated where it is needed, which is itself evidence that the score reflects real per-instance quality rather than noise. The size of the gain is sensitive to the chosen threshold, so the curve, rather than any single operating point, is the honest summary of what gating can deliver.

6. Discussion

The results support a single, clear reading. Evaluating synthetic time series is not one measurement problem but several, and the measures that address them are complementary rather than interchangeable. The disagreement of Figure 2 shows that the identity of the best generator is relative to the measure; the weak fidelity-utility relationship of Figure 3 shows that realism and usefulness are distinct goals; the divergent distributions of Figure 4 show that an average can hide the very samples that determine whether a generator is safe to use; and the low correlations of Table 3 show that the facets carry largely non-redundant information. No one of these measures is wrong, but no one of them is sufficient, and reporting any of them alone misrepresents the quality of a synthetic dataset.

The reason is structural rather than incidental. Fidelity, utility, diversity, and privacy are logically independent objectives, and a measure designed to track one of them is by construction insensitive to the others: a fidelity measure rewards statistical resemblance whether or not it is useful, a utility measure rewards task performance whether or not individual samples are realistic, and a privacy measure can be satisfied by a generator that is poor on both. Aggregation compounds the problem along an orthogonal dimension, because summarising a set by its mean necessarily discards the spread, and it is the lower tail of the per-instance distribution, not its centre, that governs the risk of consuming an individual bad sample. A scalar summary thus loses information twice over, once by privileging one facet and once by averaging within it.

What the analysis recommends in place of a scalar is an evaluation that is multi-faceted, per instance, and task-aware. Multi-faceted, because only a panel reports the several distinct facets a synthetic

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

dataset must satisfy; per instance, because only a distribution of sample-level scores exposes the tails that averages conceal; and task-aware, because the relative weight of the facets is not a universal constant but a function of the use to which the data will be put. The per-instance realism score studied above is a concrete illustration of the missing ingredient, since it can be computed before use and then employed as a gate that withholds or flags samples likely to be poor, turning evaluation from a retrospective verdict into a deployable safeguard.

These observations refine, rather than contradict, the message of the survey literature. Catalogues of evaluation measures establish that the options are many and that each has known strengths and weaknesses (Borji, 2019; Stenger et al., 2024). What a like-for-like comparison adds is evidence that the measures, applied to the same generators, actively disagree, and that the disagreement is informative rather than noise, because it reflects genuinely different facets of quality. The practical upshot is not to search for a single correct measure, which the structure of the problem suggests does not exist, but to assemble and report a panel, to read it per instance, and to weight it according to the downstream task.

Adopting a panel has costs, and acknowledging them is part of recommending it. Computing several measures requires more held-out real data, more compute, and, for the utility facet, real labels, none of which is free, and a per-instance score must be evaluated for every generated window rather than once per set. Where resources are constrained, the taxonomy offers a principled order of priority: the facet most closely tied to the intended use should be measured first, a per-instance score should be added next because it guards against the failure that aggregates cannot see, and the remaining facets can be sampled rather than computed exhaustively. The point is not that every measure must always be reported, but that the facets deliberately omitted should be named, so that a reader knows which failure modes a given evaluation has not ruled out rather than being left to assume that one number speaks for all of them.

7. Theoretical and Practical Implications

Theoretically, the study reframes the quality of synthetic data as an irreducibly multi-objective property. If fidelity, diversity, utility, and privacy are distinct and partly conflicting objectives, then no scalar measure can be sufficient, in the same sense that no single number can summarise a multi-criteria optimisation without a value judgement about trade-offs. The natural formal object is therefore a profile or a frontier rather than a point, and progress on evaluation consists in characterising that frontier, including the tension between privacy and the other facets, more accurately, rather than in compressing it. Per-instance granularity adds a second theoretical axis, distinguishing the quality of a distribution from the quality of its members, which an expectation cannot recover.

Practically, the recommendation for anyone producing or consuming synthetic time series is to report a panel of measures rather than a single figure, and to choose the facet weighting from the intended use. When synthetic data are generated to augment a training set, downstream utility and diversity should dominate, since the goal is to add useful and varied examples. When they are generated to be shared in place of sensitive records, privacy and fidelity must be assessed jointly, because a release that is private but useless, or useful but identifying, fails in either direction. When they are generated to stress-test a system against rare conditions, coverage of the tails of the distribution matters more than average realism. In each case a per-instance score should accompany the aggregates so that individual unsuitable samples can be filtered before they propagate.

There are governance and responsible-use implications as well. Synthetic data are increasingly

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

proposed as a mechanism for sharing data across institutional and regulatory boundaries, and a single-number quality claim is a poor basis for the trust such sharing requires. An evaluation that reports fidelity, utility, diversity, and privacy together, and that can certify individual samples before release, gives data stewards a defensible and auditable account of what a synthetic dataset does and does not preserve. Framing evaluation in these decision-aware terms, rather than as a leaderboard of one statistic, is the responsible counterpart to the technical argument advanced here.

These recommendations can be made concrete as a reporting template. An evaluation report for a synthetic time-series release should state, at a minimum, the intended use and the facet weighting it implies; the distributional fidelity in time, frequency, and feature space; the downstream utility on the relevant task with the task specified; a coverage or diversity figure; a privacy assessment whenever the data substitute for sensitive records; and the full distribution, not merely the mean, of a per-instance quality score, together with any gating threshold applied before release (Dankar & Ibrahim, 2021; van Breugel et al., 2021). Such a report is short, auditable, and far more informative than a single headline number, and it makes the trade-offs a generator has struck legible to a downstream user who did not train it. Standardising this template across a field would let releases be compared on common terms and would discourage the selective reporting of whichever measure flatters a given model.

8. Limitations and Future Research

Several limitations bound these conclusions. The comparison is illustrative: the generators, the feature encoder, and the numerical values are intended to expose the relationships among measures rather than to benchmark any particular system, and the precise figures would shift with other choices. The distributional and diversity measures depend on the feature space in which they are computed, so a different encoder could alter the fidelity rankings without changing the qualitative lesson that fidelity and utility diverge. The per-instance realism score relies on proximity to the real data manifold as a stand-in for quality, which is reasonable but not a ground truth, and defining a principled per-instance reference notion of quality remains an open problem. Finally, the study considers one application domain, and the relative importance of the facets will differ elsewhere.

These limitations indicate a productive research agenda. The most pressing need is for standardised per-instance, ex-ante, and task-aware measures for time series, together with benchmark suites that evaluate the measures themselves rather than only the generators, so that the community can compare evaluation protocols on common ground. The privacy-utility frontier deserves dedicated treatment, since the two facets trade off in ways that a panel can expose but not resolve. A meta-evaluation that asks which measures best predict real downstream outcomes would put the choice of panel on an empirical footing. And extending the per-instance gating analysis into an end-to-end pipeline, in which low-scoring samples are filtered automatically and the effect on real deployments is measured, would test whether the per-instance signal delivers its promised safeguard in practice.

Two further caveats temper the conclusions. The relative importance of the facets, and even the direction of some trade-offs, may differ across modalities, so the specific orderings observed here should not be transported uncritically from physiological sequences to, say, financial or text-like data, where prior surveys report different sensitivities (Iqbal & Qureshi, 2022). And the rank correlations of Table 3 are estimated from a finite set of generators and checkpoints, so their precise values carry sampling uncertainty and should be read as indicating broad independence among facets rather than as exact population quantities. Likewise, the gain from gating depends on the threshold and on the realism score's quality, and a poorly chosen score could gate away good samples; the curve reported above is the safeguard against reading too much into any single operating point.

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833

9. Conclusion

This article examined how the quality of synthetic time series should be measured, treating the evaluation measures rather than the generators as the primary objects of study. Assembling a taxonomy of distributional, utility, diversity, per-instance, and privacy measures and applying a panel of them to several generators on equal footing, it found a consistent pattern: the measures disagree on which generator is best, distributional fidelity is only weakly related to downstream utility, and aggregate scores conceal a tail of low-quality individual samples. Each measure is informative, but none is sufficient, and the disagreements among them are themselves evidence that they capture different and largely non-redundant facets of quality.

The practical conclusion is that evaluating synthetic time series calls for a multi-faceted, per instance, and task-aware protocol rather than any single score. Producers and consumers of synthetic data should report a panel of measures, inspect per-instance distributions rather than only their means, and weight the facets according to the intended use, whether augmentation, sharing, or stress-testing. For the field of data science and big data technology, where synthetic data are increasingly relied upon to train models, share sensitive information, and probe rare conditions, adopting such evaluation is a prerequisite for using these data responsibly and for trusting the decisions built upon them.

Acknowledgement

The authors thank colleagues working on synthetic data generation and evaluation for helpful discussions that informed the framing of this comparative study.

References

Alaa, A., van Breugel, B., Saveliev, E. S., & van der Schaar, M. (2022). How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In Proceedings of the 39th International Conference on Machine Learning (pp. 290–306). PMLR.

Borji, A. (2019). Pros and cons of GAN evaluation measures. Computer Vision and Image Understanding, 179, 41–65. https://doi.org/10.1016/j.cviu.2018.10.009

Borji, A. (2022). Pros and cons of GAN evaluation measures: New developments. Computer Vision and Image Understanding, 215, 103329. https://doi.org/10.1016/j.cviu.2021.103329

Brophy, E., Wang, Z., She, Q., & Ward, T. (2023). Generative adversarial networks in time series: A systematic literature review. ACM Computing Surveys, 55(10), 1–31. https://doi.org/10.1145/3559540

Dankar, F. K., & Ibrahim, M. (2021). Fake it till you make it: Guidelines for effective synthetic data generation. Applied Sciences, 11(5), 2158. https://doi.org/10.3390/app11052158

Esteban, C., Hyland, S. L., & Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional GANs. arXiv preprint arXiv:1706.02633.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (Vol. 27, pp. 2672–2680).

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. Journal of Machine Learning Research, 13(25), 723–773.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6626–6637).

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (Vol. 33, pp. 6840–6851).

Iqbal, T., & Qureshi, S. (2022). The survey: Text generation models in deep learning. Journal of King Saud University - Computer and Information Sciences, 34(6), 2515–2528. https://doi.org/10.1016/j.jksuci.2020.04.001

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: A review. Data Mining and Knowledge Discovery, 33(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Jordon, J., Yoon, J., & van der Schaar, M. (2019). PATE-GAN: Generating synthetic data with differential privacy

guarantees. In International Conference on Learning Representations.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In International Conference on Learning Representations.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., & Aila, T. (2019). Improved precision and recall metric for assessing generative models. In Advances in Neural Information Processing Systems (Vol. 32, pp. 3927–3936).

Lopez-Paz, D., & Oquab, M. (2017). Revisiting classifier two-sample tests. In International Conference on Learning Representations.

Lu, W., Lu, Y., Li, J., Sigov, A., Ratkin, L., & Ivanov, L. A. (2024). Quantum machine learning: Classifications, challenges, and solutions. Journal of Industrial Information Integration, 42, 100736. https://doi.org/10.1016/j.jii.2024.100736

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O. (2018). Are GANs created equal? A large-scale study. In Advances in Neural Information Processing Systems (Vol. 31, pp. 700–709).

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., & Yoo, J. (2020). Reliable fidelity and diversity metrics for generative models. In Proceedings of the 37th International Conference on Machine Learning (pp. 7176–7185). PMLR.

Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., & Gelly, S. (2018). Assessing generative models via precision and recall. In Advances in Neural Information Processing Systems (Vol. 31, pp. 5228–5237).

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. In Advances in Neural Information Processing Systems (Vol. 29, pp. 2234–2242).

Stenger, M., Leppich, R., Foster, I., Kounev, S., & Bauer, A. (2024). Evaluation is key: A survey on evaluation measures for synthetic time series. Journal of Big Data, 11(1), 66. https://doi.org/10.1186/s40537-024-00924-7

Theis, L., van den Oord, A., & Bethge, M. (2016). A note on the evaluation of generative models. In International Conference on Learning Representations.

van Breugel, B., Kyono, T., Berrevoets, J., & van der Schaar, M. (2021). DECAF: Generating fair synthetic data using causally-aware generative networks. In Advances in Neural Information Processing Systems (Vol. 34, pp. 22221–22233).

Wang, Z., She, Q., & Ward, T. E. (2021). Generative adversarial networks in computer vision: A survey and taxonomy. ACM Computing Surveys, 54(2), 1–38. https://doi.org/10.1145/3439723

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In Advances in Neural Information Processing Systems (Vol. 32, pp. 7335–7345).

Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series generative adversarial networks. In Advances in Neural Information Processing Systems (Vol. 32, pp. 5508–5518).

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

ISSN: 3070-345X © 2025 INATGI (Institute of Advanced Technology and Green Innovation)
https://doi.org/10.63646/LVFT3833