SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 5 | May 2026 - Peer Reviewed Journal

A GENERATIVE ADVERSARIAL NETWORK APPROACH TO SYNTHETIC TABULAR DATA GENERATION: ARCHITECTURES, MATHEMATICAL FOUNDATIONS, AND EVALUATION PRACTICES

Mukund Kumar Singh[1], Prof. Dr. Mrs. Shivani A. Budhkar[2]
[1]*Master of Computer Applications (MCA), Progressive Education Society's Modern College of Engineering, Pune, India*
[2]*Guide, Progressive Education Society's Modern College of Engineering, Pune, India*

Article DOI: https://doi.org/10.36713/epra27893
DOI No: 10.36713/epra27893

ABSTRACT

Defence organisations collect operational, medical, cyber, logistics, and maintenance records that are valuable for model development but difficult to share because they may contain sensitive or mission-revealing information. Synthetic tabular data offers a practical way to support experimentation, benchmarking, and training while reducing direct exposure of original records. This paper presents an original review of Generative Adversarial Network (GAN)-based approaches for tabular data synthesis, with emphasis on the requirements of defenceoriented workflows. It explains the adversarial objective, the Wasserstein formulation with gradient penalty, and preprocessing methods that allow neural generators to handle numerical and categorical fields. The paper also discusses evaluation through fidelity, downstream utility, robustness, fairness, and privacy testing. Rather than treating synthetic data as automatically safe, the analysis argues for a documented validation pipeline that measures both model performance and disclosure risk before synthetic records are released or used in operational decision support.

INDEX TERMS—*Generative Adversarial Networks, Synthetic Data, Tabular Data, Defence Analytics, Privacy, WGAN-GP*

I. INTRODUCTION

Modern defence analytics depends on data collected from heterogeneous sources such as sensor logs, equipment maintenance systems, cyber-security alerts, personnel records, and field medical reports. These datasets can improve forecasting, anomaly detection, resource allocation, and decision support, yet they are rarely easy to share. Access restrictions are often necessary because a single table may reveal personal information, operational tempo, asset readiness, or classified patterns of activity. Synthetic data generation is therefore attractive: it can create artificial records that retain useful statistical structure while avoiding a direct copy of the original rows.

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [1], learn to generate samples through competition between two neural networks. Their early impact was most visible in image generation, but many defence datasets are tabular rather than visual. Tabular synthesis is difficult because columns may combine continuous measurements, categorical codes, time-derived indicators, missing values, and skewed distributions. A useful synthetic table must reproduce relationships between columns without memorising sensitive examples.

This paper focuses on GAN-based tabular synthesis as a controlled data-enabling method rather than as a replacement for governance. The discussion brings together the mathematical basis of GANs, stabilisation methods such as Wasserstein GANs and gradient penalty, tabular-specific design choices such as mode-specific normalisation, and evaluation methods such as Train on Synthetic, Test on Real (TSTR). The central argument is that synthetic defence data should be judged through a privacy–utility–fidelity balance: it must be realistic enough to support analysis, useful enough for downstream tasks, and safe enough to resist disclosure attacks.

II. LITERATURE REVIEW

Early synthetic-data techniques often relied on parametric distributions, copulas, bootstrapping, or rule-based simulation. These methods remain useful when the data-generating process is well understood, but they can struggle with complex dependencies among mixed-type variables. Deep generative models were introduced to learn richer distributions directly from data, and GANs became a major approach because they do not require explicit likelihood estimation.

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 5 | May 2026 - Peer Reviewed Journal

A. Adversarial Generative Models

A standard GAN contains a generator (G), which maps random noise to a candidate sample, and a discriminator (D), which attempts to separate generated samples from real samples. The original formulation optimises a minimax game [1]. In practice, this game can be unstable: the discriminator may become too strong, gradients may vanish, or the generator may collapse to a narrow set of outputs. These weaknesses are especially problematic for tabular data, where rare categories and minority patterns may be operationally important.

The Wasserstein GAN (WGAN) reframed the objective by using the Earth-Mover distance between real and generated distributions [2]. This distance gives a smoother learning signal when the two distributions have little overlap. Gulrajani et al. later introduced WGAN with gradient penalty (WGAN-GP), replacing weight clipping with a penalty that encourages the critic to satisfy the Lipschitz constraint [3]. These improvements are widely used when training must be stable across diverse data distributions.

B. GANs for Tabular Data

Tabular data differs from images because its columns do not share a uniform spatial structure. Some fields are numerical, some are categorical, and others encode dates, ranks, or identifiers. Conditional GANs first demonstrated how class or auxiliary information could guide generation [7]. Building on this idea, the Conditional Tabular GAN (CTGAN) addressed mixed-type tables by combining conditional sampling with mode-specific normalisation for continuous variables [4]. Its training-by-sampling strategy also helps expose the model to underrepresented categories, which is important for defence scenarios where rare events may carry high consequence.

Earlier tabular GAN work such as Table-GAN demonstrated that adversarial models could synthesise database-style tables while considering information leakage [8]. Other tools and frameworks, including the Synthetic Data Vault (SDV), provide practical pipelines for modelling, sampling, and evaluating synthetic relational or tabular data [5]. More recent work explores graph-based, autoencoder-based, transformer-based, and diffusion-based approaches for tabular synthesis. These methods broaden the design space, but GANs remain relevant because they offer efficient sampling and a mature body of research on stability, privacy attacks, and evaluation.

C. Defence-Oriented Requirements

Defence use cases impose constraints that are stronger than those found in many open research benchmarks. First, rare classes such as equipment failures, intrusion attempts, or emergency medical outcomes must not disappear during generation. Second, synthetic records must avoid revealing identities, locations, schedules, or operational patterns. Third, the generation process should be documented so that analysts can explain how synthetic data was produced and where it should not be used. These constraints make evaluation as important as model design.

TABLE I
DESIGN PRIORITIES FOR DEFENCE SYNTHETIC TABLES

Priority	Implication for GAN Design
Rare events	Conditional sampling and class-aware evaluation
Mixed data types	Separate handling of numerical and categorical fields
Operational security	Privacy testing before release
Auditability	Documented preprocessing and model settings
Task relevance	TSTR or task-specific validation

III. METHODOLOGY

A robust synthetic-data workflow should define the modelling objective, transform raw records into a learnable representation, train the generator, and evaluate the outputs before use. The following subsections summarise these stages for tabular GANs.

A. Adversarial Objective

Let $x \sim P_{data}$ denote real records and $z \sim P_z$ denote samples from a latent distribution such as a standard normal distribution. The generator produces $G(z)$, while the discriminator estimates whether an input came from the real dataset.

The classical GAN objective is

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim Pdata}[\log D(x)] + \mathrm{E}_{z\sim Pz}[\log(1 - D(G(z)))]. \quad (1)$$

At equilibrium, the generated distribution should approximate the real distribution. For defence data, however, approximation alone is not enough; the generator should also preserve taskrelevant patterns while avoiding record-level memorisation.

This adversarial relationship is shown conceptually in Figure 1.

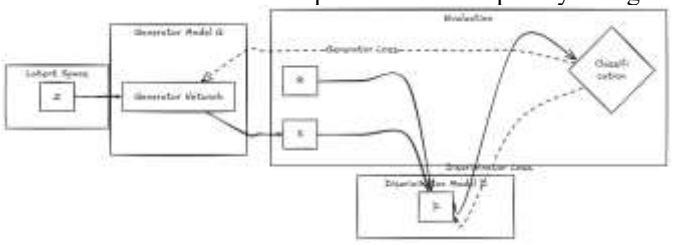


Fig. 1. Conceptual GAN workflow in which latent noise is transformed by a generator and assessed by a discriminator or critic.

B. Wasserstein Training with Gradient Penalty

WGAN replaces the discriminator with a critic that assigns scalar scores rather than probabilities. The critic is trained to assign higher scores to real samples and lower scores to generated samples. The WGAN-GP critic loss can be written as

$$L_D = \mathbb{E}_{\tilde{x}\sim P_g}[D(\tilde{x})] - \mathbb{E}_{x\sim P_r}[D(x)] + \lambda \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}\left[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2\right], \quad (2)$$

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 5 | May 2026 - Peer Reviewed Journal

where $x\tilde{}$ is generated data, x is real data, $x\hat{} = \epsilon x + (1 - \epsilon)x\tilde{}$ is an interpolated sample, and $\epsilon \sim U(0,1)$. The coefficient λ controls the strength of the gradient penalty. This formulation encourages a smoother critic and usually provides more reliable gradients than the original GAN loss.

C. Preprocessing for Mixed-Type Tables

The raw table should be transformed before training. Numerical fields may require clipping, missing-value treatment, scaling, or mode-specific normalisation. Categorical fields are commonly represented through one-hot vectors or embeddings. Identifiers that have no modelling value, such as names, record numbers, and exact device IDs, should normally be removed or replaced by lower-risk derived attributes.

1) Mode-Specific Normalisation: Continuous columns in operational datasets are often skewed or multimodal. CTGAN models each continuous column with a variational Gaussian mixture and represents a value by a normalised scalar plus a mode indicator [4]. This representation helps the generator learn different regimes, such as routine operating conditions and high-stress operating conditions, without forcing a single Gaussian assumption.

2) Categorical Attributes: Categorical attributes require careful treatment because rare categories may be washed out by standard minibatch sampling. Conditional vectors can force the generator to produce records from selected categories during training. This is useful when the analyst must preserve low-frequency but important events such as mission aborts, cyber incidents, or rare maintenance faults.

TABLE II
COMMON TRANSFORMATIONS FOR TABULAR GAN INPUTS

Field Type	Typical Transformation	Output Layer
Bounded numerical	Min–max or robust scaling	Tanh/linear
Skewed numerical	Quantile or log transform	Tanh/linear
Multimodal numerical	Mixture mode plus scalar	Tanh + softmax
Categorical	One-hot vector or embedding	Softmax
Sensitive identifier	Remove, hash, or generalise	Not generated directly

D. Network Architecture and Training Loop

For tabular data, the generator and critic are usually multilayer perceptrons. The generator maps a latent vector, optionally concatenated with a condition vector, to the transformed feature space. The critic receives transformed real or synthetic records and returns a scalar score. Leaky ReLU activations are common in the critic, while the generator output is partitioned according to the feature schema. In WGAN-GP, batch normalisation in the critic is typically avoided because it mixes samples within a batch and can interfere with the gradient penalty.

The training loop alternates between critic and generator updates. The critic is updated several times for each generator update by comparing real batches, generated batches, and interpolated batches used for the gradient penalty. After training, generated samples are inverse-transformed back to the original schema and screened for invalid values, duplicate records, and privacy risks. Figure 2 summarises this pipeline.

IV. EVALUATION FRAMEWORK

Synthetic data should not be accepted solely because it appears realistic. A defence-oriented evaluation must consider statistical fidelity, model utility, privacy, and operational suitability.

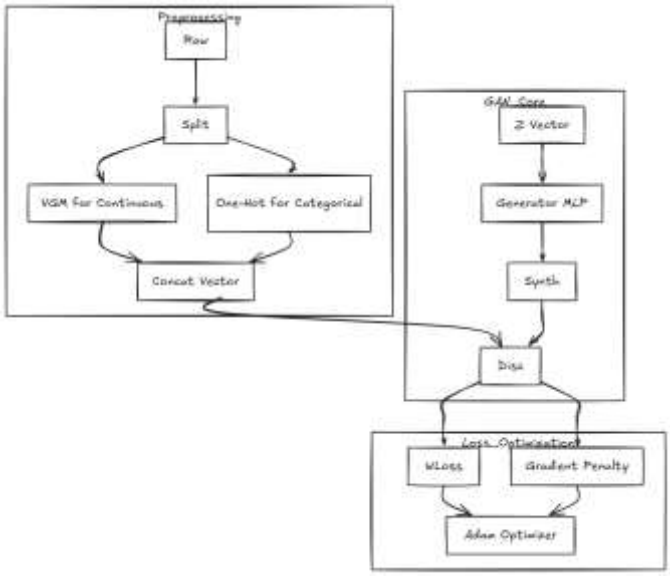


Fig. 2. Synthetic tabular-data pipeline covering preprocessing, adversarial training, inverse transformation, and evaluation.

A. Fidelity Assessment

Fidelity measures whether synthetic columns and relationships resemble the real data. Univariate tests, such as distributional comparisons for numerical fields and frequency comparisons for categorical fields, provide a first check. Pairwise correlations, contingency tables, mutual information, and domain constraints provide stronger evidence that relationships between fields have been preserved. However, high fidelity is not automatically desirable if it comes from memorising original records.

B. Utility Assessment

The Train on Synthetic, Test on Real (TSTR) protocol trains a downstream model on synthetic data and tests it on held-out real data. This approach is useful because it evaluates whether synthetic data preserves patterns that matter for a task, such as fault prediction or alert prioritisation. A complementary Train on Real, Test on Synthetic (TRTS) test can reveal whether the synthetic set covers the decision regions learned from real data.

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 5 | May 2026 - Peer Reviewed Journal

Utility should be reported with uncertainty, because a single train-test split can hide instability.

C. Privacy and Disclosure Assessment

Synthetic data can still leak information. If a generator overfits, synthetic records may be close to real training records, enabling membership inference or attribute inference attacks. Prior work on synthetic health data shows that privacy risk must be measured empirically rather than assumed away [6]. When stronger guarantees are required, differentially private training methods and private aggregation approaches, including DP-SGD and PATE-GAN, provide relevant design options for bounding disclosure risk [9], [10]. Defence datasets require an even more conservative approach: nearest-neighbour distance checks, duplicate detection, membership inference simulations, and expert review should be completed before release.

TABLE III
EVALUATION CHECKLIST BEFORE SYNTHETIC DATA RELEASE

Dimension	Recommended Evidence
Fidelity	Distribution, correlation, and constraint checks
Utility	TSTR/TRTS performance on mission-relevant tasks
Privacy	Distance, duplicate, and membership-inference tests
Fairness	Subgroup utility and error-rate comparison
Governance	Documented source data, preprocessing, and limits

V. DISCUSSION

The main risk in synthetic-data projects is treating a generated table as both realistic and safe without testing those claims. In practice, fidelity, utility, and privacy can conflict. A model that preserves every rare pattern may be valuable for training but may also reveal sensitive outliers. A model with strong noise may protect privacy but fail to support downstream prediction. The appropriate balance depends on the intended use, the sensitivity of the source data, and the consequences of analytical error.

For defence applications, synthetic data is best viewed as an intermediate artefact for controlled experimentation. It can help researchers develop code, compare algorithms, train analysts, and share low-risk examples with partners. It should not be used to make operational decisions unless its generation process, evaluation evidence, and limitations have been reviewed. Human oversight remains necessary because some risks, such as revealing a deployment pattern through a combination of fields, may not be detected by automated metrics alone.

VI. CONCLUSION

GANs provide a flexible framework for generating synthetic tabular data, and stabilised variants such as WGANGP make adversarial training more practical. Tabular-specific techniques, especially conditional sampling and mode-specific normalisation, help the generator handle mixed data types and rare categories. These strengths make GAN-based synthesis relevant to defence analytics, where useful data is often sensitive and difficult to share.

The paper's key conclusion is that synthetic data should be governed through evidence rather than assumption. A responsible workflow should remove direct identifiers, transform mixed-type features carefully, train with stability controls, and evaluate the resulting records across fidelity, utility, privacy, fairness, and governance dimensions. When this process is followed, synthetic data can support research and collaboration while reducing the exposure of sensitive defence records.

REFERENCES

1. *I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems, vol. 27, 2014.*
2. *M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," arXiv preprint arXiv:1701.07875, 2017. [Online]. Available: https://arxiv.org/ abs/1701.07875*
3. *I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in Advances in Neural Information Processing Systems, vol. 30, 2017. [Online]. Available: https://arxiv.org/abs/1704.00028*
4. *L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in Advances in Neural Information Processing Systems, vol. 32, 2019. [Online]. Available: https://arxiv.org/abs/1907.00503*
5. *N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016, pp. 399–410.*
6. *C. Chen et al., "Membership inference attacks against synthetic health data," Journal of Biomedical Informatics, vol. 125, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8766950/*
7. *M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014. [Online]. Available: https://arxiv. org/abs/1411.1784*
8. *N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," Proceedings of the VLDB Endowment, vol. 11, no. 10, pp. 1071–1083, 2018. [Online]. Available: https://www.vldb.org/pvldb/vol11/p1071-park.pdf*
9. *M. Abadi et al., "Deep learning with differential privacy," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 308–318. [Online]. Available: https: //arxiv.org/abs/1607.00133*
10. *J. Jordon, J. Yoon, and M. van der Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," in International Conference on Learning Representations, 2019. [Online]. Available: https://openreview.net/pdf?id=S1zk9iRqF7*