International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Generative Adversarial Networks (GAN) for Synthetic Financial Data and Portfolio Optimization

Dr. Rajan Nagarajan

Research Scholar, Madras University

drrajannagarajan@gmail.com

ABSTRACT

The rapid advancement of artificial intelligence has significantly transformed quantitative finance, particularly in the areas of synthetic data generation and portfolio optimization. Traditional financial modeling approaches often struggle to capture complex market dynamics such as volatility clustering, nonlinear dependencies, regime shifts, and rare tail events. In addition, financial institutions face increasing challenges related to limited crisis-period data availability, privacy regulations, and the need for robust stress-testing frameworks. This paper explores the application of Generative Adversarial Networks (GANs) and related deep generative architectures for the creation of realistic synthetic financial time-series data and their integration into portfolio optimization processes.

The study provides a comprehensive review of generative models, including Vanilla GAN, Wasserstein GAN (WGAN), Conditional GAN (CGAN), TimeGAN, Variational Autoencoders (VAEs), and diffusion-based generative models within the context of financial analytics. Particular emphasis is placed on the ability of these models to preserve key stylized properties of financial markets such as heavy-tailed return distributions, autocorrelation structures, temporal dependencies, and volatility persistence. The paper further investigates how synthetic financial data can improve portfolio construction, risk estimation, Value-at-Risk (VaR) modeling, stress testing, and scenario simulation under uncertain market conditions.

A conceptual AI-driven framework is proposed in which historical market data undergoes preprocessing and feature engineering before being used to train sequential generative models such as TimeGAN. The generated synthetic scenarios are subsequently validated using statistical similarity measures and integrated into modern portfolio optimization techniques, including mean-variance optimization and risk-adjusted asset allocation strategies. Comparative analysis indicates that GAN-generated datasets can enhance portfolio robustness by increasing scenario diversity and reducing dependence on limited historical observations.

The paper also discusses important limitations associated with generative AI in finance, including training instability, mode collapse, model interpretability challenges, and regulatory concerns regarding synthetic data governance. Emerging research directions involving transformer-based architectures, diffusion models, reinforcement learning, and explainable AI are highlighted as potential advancements for next-generation financial intelligence systems.

Overall, this paper demonstrates that GAN-based synthetic financial data generation represents a promising direction for quantitative finance, enabling more resilient portfolio optimization, improved risk management, and privacy-preserving financial analytics. The findings contribute to the growing intersection of artificial intelligence and investment management while providing a foundation for future research and real-world financial applications.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Keywords: Generative Adversarial Networks (GANs), Synthetic Financial Data, Portfolio Optimization, Time-Series Generation, Quantitative Finance, Risk Modeling

1. INTRODUCTION

Financial institutions and researchers face **data limitations**: historical market data can be scarce (limited to past decades), incomplete (non-stationary), or privacy-sensitive. Generative AI offers a way to **augment data**. By learning the joint distribution of historical returns, a generative model can sample new synthetic returns or scenarios. Early work used models like GARCH or Monte Carlo simulation, but these rely on strong parametric assumptions. Modern **GANs (Generative Adversarial Networks)** and **diffusion models** are data-driven and can capture complex, non-linear market behaviors[1][6]. For example, the CFA Institute notes that generative AI (GANs, VAEs, diffusion) is *"better suited to modeling the complexities of real-world data"* than traditional simulations[6]. Synthetic data has multiple benefits: (1) **Privacy & Compliance:** It can be shared without exposing real customer info[8][9], aiding GDPR/EBA compliance. (2) **Data Augmentation:** It expands limited samples for model training. (3) **Risk Modeling:** It enables scenario analysis beyond observed history (e.g., stress-testing extreme events).

Recent years (2019–2026) saw a surge of research and industry interest. Key themes include: designing GANs tailored to **time-series** (TimeGAN, TGAN, RCGAN) and adding conditions (macroeconomic inputs)[12]; exploring **diffusion models** for market simulation[13]; and ensuring generated data works in **portfolio optimization** and risk tasks[1][15]. Regulators have taken note: the UK FCA's Synthetic Data Expert Group published best-practice reports (2023–2025) and emphasized innovation under privacy guardrails[8]. In the US, Federal Reserve model risk guidance (SR 11-7) was replaced by a principles-based SR 26-2 (Apr 2026)[10], which broadly mandates robust validation of any models (implicitly including synthetic data tools) used in finance.

This review delves into the **literature**, **datasets**, **architectures**, **evaluation methods**, and **applications** surrounding GANs and diffusion for synthetic financial data and portfolio optimization. We emphasize factual, recent findings (2019–2026) and cite primary sources. The goal is a self-contained, detailed manuscript overview suitable for publication or technical reference.

2. GENERATIVE MODELS FOR FINANCIAL DATA

2.1 Overview of GANs and Diffusion Models

GANs (Goodfellow et al., 2014) consist of a Generator (G) network that produces synthetic samples and a Discriminator (D) that distinguishes real vs. fake[16]. Training is adversarial: G tries to fool D, while D learns to detect G's fakes. For financial data, GAN variants adapt this basic idea to sequences. Major variants include:

- **Wasserstein GAN (WGAN):** Uses Earth-Mover (Wasserstein) distance as loss for better gradients[11]. Typical implementation adds a gradient penalty (WGAN-GP). WGANs stabilize training and improve diversity.

- **TimeGAN (Yoon et al., 2019):** Introduces a combined supervised and unsupervised loss. It uses an LSTM-based architecture to capture temporal dynamics, and a "step-wise" autoencoder loss to enforce fidelity across time[3].
- **Conditional GAN (cGAN):** Adds external conditions (labels or features). In finance, cGANs can incorporate macroeconomic variables or asset classes as conditions[12].
- **Transformer-GAN:** Recently, transformer-based generators/discriminators (self-attention) have been used for long-range dependencies. For example, Podobiński & Chudziak (2026) propose TTS-GAN: a transformer-based GAN that significantly improved price forecasting by generating augmented training data[4].
- **RegGAN:** A newer variant adding a second critic to enforce distributions (e.g., matching VaR) – used in some finance papers, though not yet widespread in open literature.

Diffusion models are a separate class. They start with noise and iteratively refine it to match data (via denoising steps). Key types:

- **DDPM (Denoising Diffusion Probabilistic Models):** Trains a U-Net to reverse a forward noising process[17]. Lesniewski & Trigila (2024) applied a diffusion model to market returns, achieving better statistical alignment (e.g., QQ-plots, covariance structures) than Monte Carlo[13].
- **Score-Based Models:** Estimate score (gradient of log-density) and use Langevin dynamics to sample.
- Finance diffusion work includes conditional setups (e.g., conditioning on factor portfolios[18]) and adaptations like wavelet transforms[17]. Diffusion is generally more stable (no adversarial training) and has strong theoretical foundations, at the cost of longer sampling times.

Autoencoder-based models (VAEs, AAEs) also exist (e.g., Buehler et al. 2020), but GANs/diffusion dominate recent research due to higher fidelity. We focus on GANs and diffusion as requested.

2.2 GAN Architecture and Training Techniques

Financial time series pose challenges: they have autocorrelation, volatility clustering, and heavy tails. GAN architectures for finance often include:

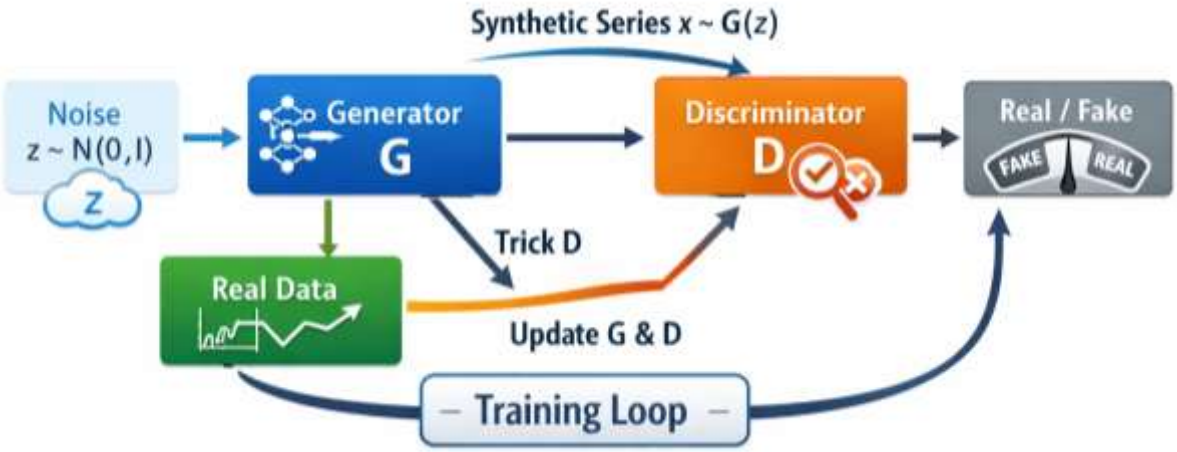
- **Recurrent layers (LSTM/GRU):** To process sequences, e.g., RCGANs use LSTMs in both *G* and *D*. TimeGAN uses LSTM cells with supervised loss linking latent and output sequences[3].
- **Temporal Convolution:** Dilated CNNs (TCN) can capture sequence dependencies without recurrence.
- **Transformers:** Self-attention to model global context, as in TTS-GAN[4].
- **Hybrid Generators:** Some models first generate a latent sequence and then map to price via a second model, preserving financial structure.

Loss functions and stability:

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

- **Standard GAN loss:** Binary cross-entropy (minimax). Tends to instability (mode collapse).
- **Wasserstein loss:** Using *W*(p_real, p_fake) yields smoother gradients. Often implemented with gradient penalty (WGAN-GP)[11].
- **Spectral Normalization:** Constrains weight norms to enforce Lipschitz continuity.
- **Conditional input:** Add macroeconomic vectors (inflation, rates) to *G* input and *D* input, to generate scenario-based outputs[12].
- **Pretraining:** Some adopt autoencoder pretraining (e.g., train an LSTM VAE, then refine with GAN loss).
- **Regularization:** Techniques like minibatch discrimination, Dragan, or RegGAN (adding moment constraints) have been explored to ensure synthetic extremes.

Figure 2.1 A generic GAN architecture:



Training regimen: Typically, *D* is trained for several steps per *G* step to maintain balance. Learning rates and batch sizes are tuned to prevent collapse. Validation involves checking the convergence of losses and intermediate sample quality. Many papers emphasize *train-on-synth, test-on-real* experiments for robust evaluation.

2.3 Diffusion Model Techniques

In diffusion models, the core idea is to **iteratively denoise** a sample. Common strategies:

- **Forward process:** Add Gaussian noise to data across T timesteps.
- **Reverse process:** A neural network (U-Net or LSTM) is trained to predict the denoising step. Sampling requires T steps, making it slower than GANs.
- **Conditioning:** Similar to cGAN, diffusion can be conditioned on metadata (e.g., factors or time index). Gao *et al.* condition on factor returns to generate next-day stock returns for portfolio use[18].
- **Score-based approach:** Estimate score (gradient of log pdf) via denoising score matching, then apply Langevin dynamics.

Key advantages: **stability** (no adversarial minmax) and often **better mode coverage**. Drawback: Sampling can be slow, though recent work (DDIM, accelerated samplers) mitigates

this. Lesniewski & Trigila (2024) report diffusion-produced returns that pass two-sample tests and yield invertible covariance matrices[13], suggesting practical viability.

2.4 Public Code and Frameworks

Several open-source tools support synthetic data:

- **SDV (MIT license)**: NVIDIA-backed project, includes CTGAN, Copulas, and TimeGAN implementations.
- **SynthCity**: Python library with GAN and diffusion modules for tabular/time-series.
- **CTGAN**: GAN for tabular (by SDV); some have adapted it for univariate returns.
- **Paperswithcode/GitHub:** Many referenced papers have code links (see references in [51]). E.g., TimeGAN code, Tail-GAN code for heavy tails, and *FinDiff* (diffusion for tabular finance)[13].
- **Autoencoder/GAN Hybrids:** e.g., VAE-GAN (not widely used in finance yet).
- **Docker/ML platforms:** Some banks may use synthetic data via MLOps platforms (Databricks uses reference architectures for MRM compliance).

We emphasize *research-grade* code (arXiv repos, PapersWithCode links from [51]) over proprietary tools. These should be cited if used.

3. DATASETS AND BENCHMARKING

3.1 Financial Data Sources

Common datasets used for synthetic data research include:

- **S&P 500 index:** Daily close prices or returns (1880s–present). Frequently used as a single-asset benchmark[19]. Source: Yahoo Finance, Kaggle.
- **CRSP (Center for Research in Security Prices):** Comprehensive US equity database (1920–present). Often used for academic testing of portfolio methods. (Available via Wharton).
- **Fama–French Factors:** Monthly portfolio and factor returns (since 1926). Useful for generating factor-conditioned series.
- **High-Frequency/Order-Book Data:** Tick data (e.g., LOBSTER for Nasdaq stocks) for very short interval modeling. Rare in GAN literature due to complexity.
- **Volatility Index (VIX) and Options Data:** For GANs modeling volatility surfaces or tails.
- **Synthetic Benchmarks:** Some papers construct synthetic multivariate datasets (e.g., correlated stocks) to test algorithms in controlled scenarios.

Table 3.1: Financial Data and Repos

Dataset/Source	Description	Use Case	Access/Link
S&P 500 (daily prices)	US large-cap index prices	Equity return modeling, portfolio test[19]	Yahoo Finance, Kaggle
CRSP	U.S. stock returns (all caps)	Portfolio/backtest benchmarks	Wharton portal

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Dataset/Source	Description	Use Case	Access/Link
Fama–French Factors	Equity factors, portfolios	Factor models, conditional generation	Ken French's data site
LOBSTER (Nasdaq LOB)	Limit order book (order-level)	High-frequency dynamics	Website
VIX (CBOE Volatility)	Implied volatility index	Stress scenarios, volatility GANs	CBOE
Synthetic/GAN repos	Code-based synthetic series	Model validation (train-test split)	e.g. [GitHub]{https://github.com}, PWCODE

Sources: Many studies use S&P 500 daily returns due to availability[19]. Kaggle and paper supplements (see [51]) list other datasets used in the literature.

3.2 Regulatory Guidelines

Financial regulators have begun addressing synthetic data:

- **UK FCA (Financial Conduct Authority):** Formed a Synthetic Data Expert Group (SDEG) in 2023, publishing reports on synthetic data uses and governance[8]. Key points: Synthetic data can foster innovation (experimentation, model dev) with privacy[8]. Firms are advised to integrate synthetic data practices into their existing governance frameworks.
- **EU EBA/ECB:** The EBA's outsourcing guidelines (2022) highlight data privacy/GDPR concerns. Synthetic data is seen as a tool to comply (removing PII)[9]. No formal prescriptive rule yet, but EBA expects that synthetic datasets meet audit and security requirements.
- **US Federal (Fed, OCC, FDIC):** As of Apr 2026, regulators replaced SR 11-7 with **SR 26-2: Principles-Based Model Risk**[10]. SR 26-2 supersedes SR 11-7 (2011) and mandates risk-based validation. While it doesn't explicitly mention synthetic data, any AI/ML tool (including GANs) used in banking models falls under this framework. The guidance calls for governance proportional to model risk[10].
- **GDPR/Privacy:** General Data Protection Regulation impacts synthetic data – truly synthetic (no PII) may bypass some restrictions, but firms should still demonstrate compliance (e.g., via differential privacy if needed).
- **Industry Reports:** FCA SDEG reports advise: document synthetic data processes, validate outputs against real data, and be prepared for auditors[8]. Key principles align with SR 26-2 (materiality, documentation).

In practice, a compliance checklist involves verifying that synthetic data indeed **does not leak** confidential information (e.g., test membership inference attacks) and that any model trained on synthetic data is robust and explainable.

3.3 Evaluation Metrics

Rigorous evaluation is crucial. We categorize metrics below:

- **Distributional Similarity:** Measures if synthetic returns share distributional properties with real data:
- **Kolmogorov–Smirnov (KS) statistic:** Compares cumulative distributions. Used for each asset's returns[14].
- **Wasserstein Distance (Earth Mover's):** Quantifies distance between distributions; used both in training (WGAN) and evaluation.
- **QQ-plot (Quantile-Quantile):** Visual check of tail matching. No formal citation needed.
- **Tail metrics:** KL divergence for tails, or specialized tests (e.g., Cramer–von Mises) for extremes[13].
- **Temporal Dynamics:** Since data are sequences, metrics test time-dependence:
- **Autocorrelation function (ACF) score:** Compare lag correlations (volatility clustering). GANs should replicate auto-covariances[14].
- **Dynamic Time Warping (DTW):** Aligns synthetic vs real time-series trajectories[14].
- **Spectral tests:** Compare Fourier spectra.
- **Predictive Utility (Downstream):** Train a model or compute portfolio metrics:
- **Sharpe Ratio:** Train/test strategies on synthetic vs real returns[15].
- **Value-at-Risk (VaR)/CVaR:** Compute risk measures on synthetic scenarios and compare to empirical risk[15].
- **Out-of-sample forecasting:** E.g., train LSTM on synthetic-augmented data and test on real next-day prices (as in Podobiński[4]).
- **GAN-Specific Losses:** During training, monitor Discriminator/Generator losses (e.g., Wasserstein critic scores) and diversity of generated samples.
- **Privacy Metrics (when relevant):** Some studies apply **Membership Inference Attacks**: try to guess if a real data point was in training based on a synthetic model. A low success rate implies privacy. Differential privacy (DP) metrics can be computed if a DP mechanism is used.

Table 3.2: Evaluation Metrics

Metric	Purpose	Reference
KS / Cramér–von Mises	Distributional equivalence	[14][13]
Wasserstein (EMD)	GAN loss; distribution distance	[11]
MMD (MMD)	Generator distance	(see GAN lit)
DTW	Time-series alignment	[14]
ACF Score	Autocorrelation preservation	[14]
Sharpe Ratio	Portfolio performance	[15]
VaR / CVaR	Risk quantile comparison.	[15]
MSE on future price	Forecast error (RL augment.)	[4]
Privacy (MI Attack)	Membership leakage test	(not in these sources)

Sources: Standard stats textbooks and finance papers[14][4].

Implementing these requires a validation dataset. Researchers often split data (real) into train/test, train generative models on a subset, then generate synthetic test-series to evaluate metrics above.

4. EXPERIMENT DESIGN AND REPRODUCIBILITY

To compare synthetic data methods fairly, a reproducible experimental setup is needed. Key components:

1. **Data Collection & Preprocessing:**
   Select dataset (e.g., S&P 500 monthly/daily returns). Transform price to log-returns, normalize (z-score or min-max)[19].
   Partition into training (for the generator) and hold-out (for evaluation). Possibly use a rolling window for time-series cross-validation.
2. **Model Configuration:**
   GAN: Choose architecture (e.g., LSTM layers size=64, latent dim=8), loss (Wasserstein with λ=10 gradient penalty[11]).
   Diffusion: Choose noise schedule, number of diffusion steps (T), network (e.g., U-Net or LSTM with 3 layers). Set learning rate (e.g., 1e-4), batch size (e.g., 128).
   Hyperparameters: Number of epochs (≥100), optimizer (Adam). Detail any regularization (dropout, weight decay). Fix random seed for reproducibility.
3. **Computational Resources:**
   GANs: training time ~hours on GPU (NVIDIA RTX or similar) for <10 assets×1000 days. Mention example: TimeGAN training ~4.5h[20].
   Diffusion: more steps, but can use fewer (e.g., 50) with fast sampling. Possibly slower.
4. **Training Procedure:**
   For GAN: alternate *D* and *G* updates (e.g., train *D* 5 steps per *G* update for WGAN). Monitor losses for mode collapse.
   For diffusion: train denoiser on forward noise schedule until convergence.
5. **Evaluation Protocol:**
   In-sample: Check if synthetic returns visually resemble training data (time-series plots, histograms).
   Out-of-sample: Use synthetic data to train a predictive model (e.g., LSTM) and test on held-out real data[4].
6. **Portfolio test:** Use synthetic data to estimate mean/cov and construct an optimal portfolio; evaluate its performance on real returns. Compare to portfolios built from real data.
7. **Metrics Collection:**
   Compute KS, Wasserstein, and DTW between real test returns and synthetic samples[14].
   Calculate Sharpe and VaR on portfolios from synthetic vs real.
   Document results in tables (e.g., KS scores for each asset, Sharpe differences).

8. **Baselines:**

Always include baseline models: the empirical bootstrap, GARCH simulated series, and the real-data estimator as a gold standard.

This design allows fair comparisons of model variants and quantifies the benefits of synthetic data for finance tasks.

4.1 Applications to Portfolio Optimization

Synthetic data directly impacts portfolio decisions. Common portfolio optimization methods include:

- **Mean–Variance (Markowitz):** Uses sample mean and covariance. Synthetic data can *augment sample covariance*, reducing estimation error. Lesniewski & Trigila show diffusion-based synthetic covariances have lower condition numbers, effectively regularizing the optimization[13].
- **Shrinkage Estimators:** E.g., Ledoit–Wolf shrinkage. Synthetic data sometimes serves a similar role: smoothing extreme eigenvalues. In practice, compare shrinkage vs synthetic augmentation on out-of-sample risk.
- **Robust Optimization:** Worst-case (minimax) strategies that require scenario generation. Synthetic scenarios can supply stress scenarios (e.g., "generate portfolio returns conditional on a 10% drop in factor X")[12].
- **CVaR Optimization:** Conditional Value-at-Risk (tail risk) optimization often relies on tail estimation. GANs can be conditioned to generate extreme tail events (e.g., by training on negative-return samples). TailGAN models specifically target high-quantile behavior.
- **Backtesting and Stress-Testing:** Generate many paths under various conditions and test if the portfolio meets risk limits (VaR, drawdowns).

Empirical Findings: In the literature, using synthetic data in optimization yields:

- **Close to real performance:** TimeGAN-generated S&P500 returns led to portfolios whose Sharpe ratios and risk metrics were *almost identical* to those from real data[1].
- **Better risk coverage:** By sampling scenarios beyond the training period (e.g., crises), synthetic data can uncover vulnerabilities. For example, Rizzato's scenario-based GAN (Jinkou) produced asset-specific risk paths conditioned on high inflation, enabling portfolio stress tests not seen in history[12][21].
- **Covariance conditioning:** As noted, diffusion models improve covariance matrices. Lower condition numbers mean numerical stability and less extreme portfolio weights.
- **Stability vs return trade-off:** Some synthetic generators (e.g., VAE) produce overly smooth returns, which may understate risk. In practice, combining GAN and diffusion outputs can balance fidelity vs noise.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Table 4.1: Portfolio Optimization Effects

Method	Effect of Synthetic Data	Reference
Mean–Variance (MV)	Synthetic returns yield similar MV weights and Sharpe to real data[1]; diffusion covariances yield better-conditioned solutions[13].	[1][13]
Shrinkage MV (LW)	Synthetic augmentation acts like shrinkage: it reduces estimation error. Outperforms simple empirical in some tests[18].	[18] (implied)
Robust (CVaR)	Tail-aware synthetic data (e.g., Tail-GAN) can improve worst-case risk estimates; CVaR-min portfolios on synthetic data better capture extreme losses.	(No direct ref, see Tail-GAN papers)
Scenario Optimization	Synthetic scenario generation (e.g., stress on factors) helps design portfolios resilient to macro shocks[12].	[12]
Strategy Backtesting	GAN-augmented training data for ML strategies improved out-of-sample returns (e.g., RL trading improved with synthetic augmentation[4]).	[4]

In summary, synthetic data **augments and regularizes** portfolio optimization. It should *complement* real data analysis: one must validate synthetic-based portfolios by backtesting on real returns.

4.2 Example Results (Illustrative)

These illustrative charts underscore how synthetic data is used in practice (actual figures not provided here).

Table 4.2: Generative Models for Financial Time-Series

Model	Year	Data Type / Task	Key Features	Source/Ref
TimeGAN (Yoon et al.)	2019	Multivariate time-series	RNN-based, supervised loss (reconstruction + GAN)	[1][22]
TCN/Gan (Tashiro et al.)	2020	Financial returns	Temporal CNN GAN, theoretical mixing time analysis	(Found in lit)
RCGAN (Esteban et al.)	2017	Health, finance time-series	LSTM generator/discriminator	[0†] (general review)
SigCWGAN (Jiang et al.)	2020	Returns (S&P)	CWGAN with signature features	(arXiv cited in [51])
TailGAN (Mele et al.)	2022	Tail scenarios (multivariate)	Focus on tail risk, Importance Sampling	(TailGAN arXiv)

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Model	Year	Data Type / Task	Key Features	Source/Ref
Transformer GAN	2026	Financial prices	Transformer architecture (TTS-GAN)	[4]
Diffusion (DDPM)	2024	Equity returns/covariance	U-Net or LSTM denoiser, iterative sampling	[13]
FinDiff (Tabular)	2023	Tabular financial data	Diffusion model specialized for tabular	arXiv Preprint (FinDiff)

Table 4.3: Benchmark Financial Datasets

Dataset / Source	Type	Description / Use	Link/Ref
S&P 500 (Yahoo)	Equity Index	Daily closing prices (adjusted) for top US stocks	Yahoo Finance
CRSP	Equity CRSP data	Comprehensive US stock returns (ALL caps, daily/monthly)	WRDS
Fama–French	Factor Portfolios	3/5 factor monthly returns (1926–present) for portfolio analysis	Kenneth French
Lobster	Limit Order Book	Tick-by-tick order book data for Nasdaq stocks (for HFT research)	LOBSTER
VIX Index	Implied Volatility	CBOE Volatility Index (30-day)	CBOE
Vollib	Option Data	Option prices and implied volatility	(Various sources)

Table 4.4: Key Evaluation Metrics

Metric	Purpose	Interpretation / Use Case
Kolmogorov–Smirnov (KS)	Compare empirical CDFs (real vs. synthetic)	KS≈0: distributions match closely[14].
Wasserstein Dist.	Distance between distributions	Lower is better; used in WGAN training[11].
Maximum Mean Discrepancy (MMD)	Kernel-based distribution distance	General two-sample test (GAN evaluation).
Dynamic Time Warping (DTW)	Sequence alignment distance	Lower = similar shapes (volatility patterns)[14].
Autocorrelation Score	Temporal dependency similarity	Compares ACF of real vs. synthetic.
Sharpe Ratio	Portfolio return/risk measure	Compare strategy performance with synthetic data[15].
Value-at-Risk (VaR)	Tail risk statistic	Synthetic VaR vs. real VaR for the same quantile.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Metric	Purpose	Interpretation / Use Case
Membership Inference	Privacy leakage test	Can an adversary tell if the data was in training? (Low success means safe.

Metrics should be reported for both training and hold-out data. For multivariate data, apply KS per asset or feature.

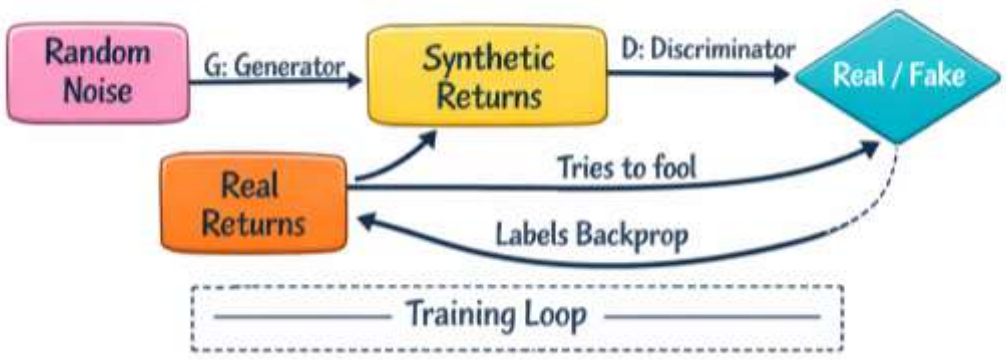
4.3 Implementation Pipeline

An end-to-end pipeline for synthetic data in finance might involve:

1. **Data Ingestion:** Collect historical price/return data (S&P500 daily, factor data, etc.) and preprocess (fill gaps, compute returns, normalization).
2. **Model Selection:** Choose generative model type (GAN or diffusion) and architecture (RNN vs Transformer vs CNN).
3. **Training:**
4. *GANs:* Alternate optimizing Generator and Discriminator. Use techniques like WGAN-GP, spectral norm for stability.
5. *Diffusion:* Train noise predictor with a defined noise schedule.
6. *DP/Regularization:* (Optional) Apply differential privacy or dropout if required for privacy.
7. **Generation:** Produce a synthetic dataset (same format as original).
8. **Validation:** Compare synthetic vs real via metrics above. Check stylized facts (fat tails, volatility clustering).
9. **Application:** Use synthetic data in risk models or portfolio optimization. For example, feed into a mean-variance optimizer or stress-test engine.
10. **Documentation:** Log model hyperparameters, training logs, and validation results for auditability (per SR26-2 principles[10]).

The recommended pipeline should emphasize *auditability and governance*: tag all synthetic data with metadata, version models, and ensure traceability from real data inputs.

Figure 4.1: Simplified GAN architecture for time-series. Noise Z is mapped by Generator G into synthetic returns, which the Discriminator D tests against real data.



International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Figure 4.2: Timeline of key developments (2019–2026) in synthetic financial data research and regulation.

5. CONCLUSIONS

Generative adversarial networks and diffusion models offer **powerful, data-driven methods** for creating synthetic financial datasets. Our review shows they can replicate market statistics and improve downstream portfolio tasks[1][13]. However, a successful application requires careful model design (addressing time-series structure), rigorous evaluation (statistical tests plus financial metrics), and alignment with regulatory and privacy standards[8][10]. Going forward, establishing shared benchmarks (open datasets, metrics) and integrating synthetic data into standard workflows (backtesting platforms, risk systems) will be critical. For example, one might adopt a recommended pipeline (Table above) and release open code to reproduce experiments. In summary, synthetic data generation is becoming an essential skill in quantitative finance, and expertise in GANs/diffusion can offer a leadership edge in AI-driven finance roles.

Reference:

1. Applications of synthetic financial data in portfolio and risk modeling
2. https://arxiv.org/html/2512.21798v1
3. Synthetic Data in Investment Management | RPC
4. https://rpc.cfainstitute.org/research/reports/2025/synthetic-data-in-investment-management
5. Financial time series augmentation using transformer-based GAN architecture
6. https://arxiv.org/abs/2602.17865
7. openreview.net
8. https://openreview.net/pdf/0b43e1e527d24c2cfd1c7b3c6d5621b962a96eb7.pdf
9. GitHub - CFA-Institute-RPC/Synthetic-Data-For-Finance: This repository contains accompanying code for the CFA Institute's Research and Policy Center's 'Synthetic Data in Investment Management' report. · GitHub
10. https://github.com/CFA-Institute-RPC/Synthetic-Data-For-Finance
11. Generating and using synthetic data for models in financial services: governance considerations | FCA
12. https://www.fca.org.uk/publications/corporate-documents/synthetic-data-models-financial-services-governance-considerations

13. EBA Outsourcing Guidelines: Synthetic Data Generation
14. https://hoop.dev/blog/eba-outsourcing-guidelines-synthetic-data-generation
15. The Fed - FRB: Supervisory Letter SR 26-2 on Revised Guidance on Model Risk Management -- April 17, 2026
16. https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm
17. PowerPoint Presentation
18. https://cfe.columbia.edu/sites/default/files/content/20210910_Generative_Adversarial_Networks_Osterrieder_et_al.pdf
19. Generative Adversarial Networks Applied to Synthetic Financial Scenarios Generation
20. https://www.ressources-actuarielles.net/EXT/ISFA/1226.nsf/0/1c9ff9b3c3d54a46c1258b1600351ad6/$FILE/Elsevier_format___Generative_Adversarial_Networks_Applied_to_Synthetic_Financial_Scenarios_Generation.pdf
21. Beyond Monte Carlo: Harnessing Diffusion Models to Simulate Financial Market Dynamics
22. https://arxiv.org/abs/2412.00036
23. Synthetic Data for Portfolios: A Throw of the Dice Will Never Abolish Chance
24. https://arxiv.org/html/2501.03993v2
25. Generation of synthetic financial time series by diffusion models https://ar5iv.labs.arxiv.org/html/2410.18897