International Journal of Informatics Engineering and Computing (IJIMATIC)
Vol. 3, No. 1, 2026
E-ISSN: 3090-9112

Geometric Structured Trend Tunneling: A Hybrid VARIMA-SVR Model for Synthetic Stock Time Series Generation

I Wayan Ordiyasa[1], Ahmad Sahal[2], Gladies Serren Kutani[3]

Abstract

This study presents a novel hybrid framework, Geometric Structured Trend Tunneling (GSTT), for generating synthetic multivariate time series data, specifically applied to stock price data of Medco Energi Internasional (MEDC), a major player in Indonesia's energy sector. The proposed model integrates the statistical power of Vector Autoregressive Integrated Moving Average (VARIMA) with the nonlinear pattern-capturing capability of Support Vector Regression (SVR), enabling high-fidelity reconstruction of temporal structures and feature dependencies in financial datasets. The dataset used spans over two decades (2003–2024) and includes core trading indicators such as Open, High, Low, and Close prices. Experimental results demonstrate that GSTT achieves excellent performance across multiple evaluation metrics, including MAE, RMSE, R^2, and KS tests, while preserving inter-feature correlations and distributional fidelity. Visual comparisons and descriptive statistics further confirm the model's ability to replicate realistic market behavior. Unlike deep generative models such as GANs or VAEs, GSTT offers a more interpretable, stable, and computationally efficient alternative for financial data augmentation, simulation, and robust AI training. This work contributes a scalable solution for addressing data scarcity in financial modeling, with potential applications in backtesting, risk analysis, and algorithmic trading simulations.

Keywords: Synthetic Time Series Stock Data Generation Hybrid VARIMA SVR Financial Simulation MEDC Indonesia

1. Introduction

Time series data plays a pivotal role in the financial market, particularly in understanding the dynamic behavior of stock prices over time.[1] Investors, analysts, and automated trading systems rely heavily on time-dependent patterns to make informed decisions.[2] However, real-world financial data often suffers from limitations such as noise, incompleteness, and data scarcity during critical periods.[3] In this context, synthetic time series generation emerges as a powerful tool to enhance data volume and diversity without compromising statistical integrity.[4] It supports robust modeling, simulation, and stress testing in volatile markets[4].

Medco Energi Internasional (MEDC), a leading Indonesian energy company listed on the Indonesia Stock Exchange, offers a compelling case study due to its strategic role in the national energy sector.[5] The volatility and structural patterns in MEDC stock data make it an ideal target for generative modeling experiments. [6] This study introduces the Geometric Structured Trend Tunneling (GSTT) algorithm, a novel method that geometrically extracts and reshapes time series trends to generate high-fidelity synthetic

Corresponding Author: I Wayan Ordiyasa, Universitas Respati Yogyakarta, Indonesia (wayanordi@respati.ac.id)
1 I Wayan Ordiyasa, Universitas Respati Yogyakarta, Indonesia
2 Ahmad Sahal, Universitas Respati Yogyakarta, Indonesia
3, Gladies Serren Kutani, Universitas Respati Yogyakarta, Indonesia

data.[7]Unlike traditional models, GSTT leverages geometric projection and structured trend manipulation to preserve temporal coherence.[8] By applying GSTT to MEDC's multivariate stock data, this research aims to bridge data gaps while maintaining the statistical and financial realism of the market.[9]

Despite the rapid advancement of synthetic time series generation methods across domains such as healthcare, energy, and finance, several critical gaps remain unaddressed. While GAN-based models have achieved notable success in forecasting financial data distributions [10] and deep learning methods have mitigated data scarcity in healthcare [11], these approaches often struggle with preserving long-range temporal structure and multivariate correlations, especially in volatile financial markets like Indonesia. Existing techniques, including diffusion models[12], copula-based frameworks [13], and hybrid GAN architectures [14], rarely integrate geometric or structure-aware mechanisms capable of explicitly modeling trend morphology over time. Additionally, challenges such as mode collapse[15], training instability, and insufficient evaluation measures for trend preservation[16] hinder the robustness of synthetic data in critical financial contexts. This research introduces the Geometric Structured Trend Tunneling (GSTT) algorithm as a response to these limitations—offering a novel framework that geometrically projects and tunnels time series trends, with a focus on generating high-fidelity synthetic data for Medco Energi Internasional (MEDC), a strategically important and underrepresented case in synthetic financial modeling.

This study aims to develop and implement a novel synthetic time series generation framework tailored for complex, multivariate financial data, with a particular focus on the stock performance of Medco Energi Internasional (MEDC). Building on prior advances in generative modeling across healthcare and finance [11]; [10], the research introduces the Geometric Structured Trend Tunneling (GSTT) algorithm to address key limitations such as mode collapse, unstable training, and inadequate preservation of temporal trends found in traditional GANs and VAEs [15]; [17]. The core objective is to generate synthetic multivariate time series data that not only mirrors the statistical fidelity of MEDC's actual market behavior but also retains its geometric and structural patterns over time. By doing so, this research contributes a structure-aware and geometry-driven method to the literature, aligning with growing interest in integrating spatial representations into temporal data synthesis [14]; [18]. Ultimately, the study seeks to expand the frontier of synthetic financial data generation for emerging markets by producing high-quality, realistic datasets that support forecasting, risk modeling, and AI-driven investment strategies.

The novelty of this research lies in its introduction of the Geometric Structured Trend Tunneling (GSTT) algorithm, a structure-aware generative framework that applies geometric transformations to multivariate trend components for synthetic time series generation. Unlike traditional models such as GANs and VAEs, which often face issues of mode collapse and temporal distortion [15]; [17], GSTT preserves the intrinsic shape and dynamics of temporal trends by operating in a geometrically projected feature space. While previous work has leveraged deep learning for synthetic healthcare and energy data [19]; [20], and incorporated spatial analysis in flood forecasting [14], few—if any—approaches have explicitly embedded geometric reasoning into financial time series synthesis. This research is among the first to apply such a method to real-world stock data from an emerging market context, focusing on Medco Energi Internasional (MEDC), thereby addressing both geographic and methodological gaps in the literature. By bridging geometric modeling and financial time series generation, this study opens new avenues for creating robust, trend-preserving synthetic datasets that can support decision-making, forecasting, and market simulation with higher temporal realism.

This research makes a significant contribution to the field of synthetic time series generation by introducing the Geometric Structured Trend Tunneling (GSTT) algorithm, which fills critical methodological and application-specific gaps in the current literature.

Unlike conventional approaches that often overlook the geometric continuity of temporal patterns [17]; [15], GSTT employs a novel geometry-driven framework to extract, transform, and regenerate multivariate financial time series with preserved trend fidelity. By applying this method to stock data from Medco Energi Internasional (MEDC), the study contributes to the underexplored intersection of geometric modeling, financial analytics, and synthetic data generation in emerging markets [10]; [14]. Furthermore, it enriches the discourse on evaluation strategies by emphasizing structural integrity and temporal coherence—dimensions often missing in existing assessment frameworks [16]. Ultimately, this research offers a robust tool for generating realistic and structurally consistent financial data, supporting more resilient forecasting models, scenario testing, and data-augmentation-driven AI applications.

2. Related Works

Recent advancements have investigated synthetic time series generation across multiple fields, including healthcare and energy, highlighting its growing relevance. In financial forecasting, Generative Adversarial Networks (GANs) have been employed to model conditional distributions of asset returns with increasing precision [21]. Within the healthcare sector, the generation of synthetic data has become essential to overcoming privacy issues and limited sample sizes, with deep learning methods taking a central role [19]. Diffusion-based models have also demonstrated the capacity to create realistic and privacy-compliant synthetic time series, particularly for electronic health records [12]. In the energy domain, data augmentation techniques have improved the handling of missing values, especially in situations where data is scarce or incomplete [20]. Researchers have applied time-frequency domain methods, such as DC-GANs, to synthesize physiological signals like ECG, further expanding the scope of synthetic time series applications [22]. Despite this progress, evaluating the quality of synthetic time series remains a challenge due to the absence of universally accepted metrics and protocols [17]. Nevertheless, the field continues to evolve, contributing meaningful innovations to smart energy systems [23] and biomedical research [24].

Studies focusing on stock market prediction—especially in the context of emerging economies—have adopted diverse computational approaches to improve accuracy and insight. Classic models such as ARIMA often fail to accommodate the non-linearities and long-term dynamics intrinsic to financial data [25]. More recent techniques involving machine learning and deep learning models like GRU and LSTM have demonstrated enhanced predictive capabilities by incorporating external variables and capturing complex data structures [26]; [27]. Generative frameworks including GANs and VAEs have also emerged as viable tools for synthesizing realistic financial data and uncovering hidden patterns in large-scale datasets [28]. However, these models still contend with notable challenges such as instability during training and vulnerability to mode collapse, which can undermine the diversity and reliability of generated data [15]. Hybrid approaches that blend traditional statistical models with advanced machine learning methods have shown promise in enhancing predictive accuracy [27]; [29]. To further improve generative model outputs, researchers have proposed novel loss functions and feature-matching techniques tailored for financial applications [10].

In parallel, researchers have begun integrating geometric and structurally informed algorithms into synthetic time series generation with the goal of maintaining temporal coherence. Numerous studies have analyzed different strategies for creating synthetic time series in areas such as healthcare and environmental monitoring, demonstrating the versatility of these methods [17]; [30]; [31]. GAN-based models in particular have proven effective in capturing complex temporal dependencies and producing high-quality synthetic sequences [32]; [33]; [14]. These frameworks have also succeeded in preserving spatial-temporal relationships, improving performance in tasks like flood prediction and dynamic

environmental modeling [14]. Tools such as GeoChron have been developed to help visualize large-scale spatial time series, facilitating interpretability and pattern discovery [18]. Moreover, the use of geometric features in AI models has enhanced the generation of structured content in artistic and visual domains, suggesting potential extensions into time series applications [34].

Additional research has addressed how synthetic data can maintain complex correlation structures among variables in financial and medical datasets. Techniques in this space range from GAN architectures [35]; [36], to statistical models based on copulas [13], and information-theoretic approaches [37]. These methods have been successfully applied to various data types, including structured tables, images, and time series [11]; [38]. Evaluation frameworks typically incorporate metrics assessing marginal distributions, feature correlations, and overall quality indicators [13]. Privacy protection remains a key concern, prompting the use of techniques such as differential privacy in synthetic data workflows [36]. The benefits of synthetic data are well recognized, offering pathways to accelerate scientific discovery, improve model generalization, and expand access to meaningful datasets without compromising sensitive information [35]; [11]. Nonetheless, ensuring that synthetic data meets rigorous standards of quality and validity remains crucial for deployment in real-world environments [39]

Finally, the evaluation of synthetic time series data has become a distinct area of focus, especially in financial and healthcare settings. Standard evaluation practices aim to quantify realism, consistency with statistical properties, and alignment with temporal trends [17]; [40]. These assessments typically combine quantitative metrics, expert judgment, and privacy verification tools [40]. Some scholars have developed integrated evaluation frameworks that address concerns such as bias mitigation, utility, fidelity, robustness, and privacy [41]; [42]. In the healthcare domain, limitations of current evaluation tools have become evident, particularly in the need for clinical validations and methods to measure temporal integrity [43]. Advanced generative architectures like VAE-GANs and GluGAN have demonstrated their utility in creating synthetic data streams for applications such as smart home energy systems and glucose tracking, emphasizing the broad applicability of synthetic data generation [44]; [33]. These innovations affirm the potential of synthetic data as a solution to pressing challenges of data scarcity, privacy, and model robustness across sectors.

3. Proposed Method

The proposed Geometric Structured Trend Tunneling (GSTT) model is built upon a hybridization of Vector Autoregressive Integrated Moving Average (VARIMA) and Support Vector Regression (SVR) to effectively capture both global and local structures in multivariate time series data. This combination leverages the statistical strength of VARIMA for trend modeling and the nonlinear learning capability of SVR for pattern refinement. $Let\ Y_t \in \mathbb{R}^n$ represent the multivariate time series at time t, where n is the number of features (e.g., Open, High, Low, Last).

Step 1: Trend Extraction via VARIMA

The VARIMA model captures linear interdependencies and trends across time and variables. A general VARIMA (p, d, q) model for the multivariate series Y_t is expressed as:

$$\Phi(B)(1 - B)^d Y_t = \Theta(B)\varepsilon_t \quad (1)$$

Where:

B is the backward shift operator: $B\ Y_t = Y_{\{t-1\}}$

$\Phi(B) = I - \Phi^1 B - \Phi^2 B^2 - \cdots - \Phi_p B^p$ is the autoregressive polynomial matrix

$\Theta(B) = I + \Theta^1 B + \Theta^2 B^2 + \cdots + \Theta_q B^q$ is the moving average polynomial matrix
d is the differencing order
ε_t is the residual error term assumed to be white noise

The VARIMA component generates a trend estimate $\hat{T}_t$, and its residuals R_t are computed as:

$$R_t = Y_t - \hat{T}_t \tag{2}$$

Step 2: Nonlinear Pattern Modeling via SVR

The residuals R_t are modeled using Support Vector Regression (SVR) to capture nonlinear structures missed by the linear VARIMA component. For each residual feature $r_t^i \in R_t$, SVR seeks a function $f_{i(x_t)}$ of the form:

$$f_{i(x_t)} = \sum_{\{j=1\}}^{N} \alpha_j^i K(x_j, x_t) + b \tag{3}$$

Where:

x_t is the input vector (e.g., lagged values of Y_t)
$K(\cdot, \cdot)$ is a kernel function (e.g., radial basis function)
α_j^i are learned coefficients from training
b is the bias term
N is the number of support vectors

SVR predicts the nonlinear correction $\hat{R}_t$, and the final generated synthetic series $\hat{Y}_t$ is reconstructed as:

$$\hat{Y}_t = \hat{T}_t + \hat{R}_t \tag{4}$$

The VARIMA layer captures structured, low-frequency trends and linear dependencies across multiple financial indicators. The SVR layer acts as a tunneling mechanism, injecting refined local variations and nonlinear residual corrections. Together, they form the Geometric Structured Trend Tunneling (GSTT) process, where time series are projected, decomposed, and tunneled through residual learning to preserve both global geometry and local dynamics.

4. Experimental Setup

Before analysis, several preprocessing steps were conducted to ensure data quality and consistency. First, the Date column was parsed and set as the time index to maintain temporal structure. Any missing values were addressed through forward filling to preserve continuity without introducing artificial trends. Then, the four numerical features (Open, High, Low, and Last) were normalized using the MinMaxScaler, scaling each value between 0 and 1 for optimal model performance. These steps helped standardize the dataset, reduce potential biases, and prepare it for effective training under the hybrid VARIMA–SVR framework.

The dataset used in this study was sourced from Investing.com, containing daily stock data for Medco Energi Internasional (MEDC) from January 2, 2003, to April 29, 2024. It consists of 5,145 records and includes five key attributes: Date, Last (Closing Price), Open, High, and Low. These attributes reflect essential market dynamics and provide a comprehensive view of MEDC's stock behavior over two decades. The dataset's temporal depth and consistency make it ideal for modeling long-term trends and testing generative algorithms. Its rich structure enables robust evaluation of the proposed hybrid VARIMA–SVR model in replicating complex financial patterns.

To quantitatively evaluate the performance of the proposed GSTT model in generating realistic multivariate synthetic time series, six statistical metrics were employed: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2 Score), Kolmogorov–Smirnov Statistic (KS Statistic), and KS P-Value. These metrics assess both the point-wise predictive accuracy and the distributional similarity between original and synthetic data.

Let y_t denote the original value at time t, and $\hat{y}_t$ denote the corresponding synthetic (predicted) value. Let T be the total number of time steps.

Mean Squared Error (MSE) measures the average squared difference between the original and synthetic values. It penalizes larger errors more than smaller ones.

$$MSE = \left(\frac{1}{T}\right) \sum_{\{t=1\}}^{T} (y_t - \hat{y}_t)^2 \tag{5}$$

Mean Absolute Error (MAE) calculates the average of the absolute differences, providing a linear perspective on the error magnitude.

$$MAE = \left(\frac{1}{T}\right) \sum_{\{t=1\}}^{T} |y_t - \hat{y}_t| \tag{6}$$

Root Mean Squared Error (RMSE) is the square root of MSE, providing an interpretable error measure in the same units as the original data.

$$\text{RMSE=sqrt}\left((1/\text{T}) \sum_{\text{t=1}}^{\text{T}} (\text{y}_\text{t}\text{-}\hat{\text{y}}_\text{t})^2 \right) \tag{7}$$

Coefficient of Determination (R^2 Score) evaluates how well the synthetic values approximate the original data. A value close to 1 indicates high accuracy.

$$R^2 = 1 - \frac{\sum_{t=1}^{T} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2} \tag{8}$$

Where $\bar{y}$ is the mean of the original values:

$$\bar{y} = \left(\frac{1}{T}\right) \sum_{\{t=1\}}^{T} y_t \tag{9}$$

Kolmogorov–Smirnov Statistic (KS Statistic) measures the maximum distance between the empirical cumulative distribution functions (ECDFs) of the original and synthetic data:

$$D = \sup_x \left| F_{n(x)} - G_{m(x)} \right| \tag{10}$$

Where:

$\text{F}_\text{n}\,(\text{x})$ is the ECDF of original data

$\text{G}_\text{m}\,(\text{x})$is the ECDF of synthetic data.

KS P-Value quantifies the statistical significance of the similarity between the original and synthetic distributions. A high p-value (typically > 0.05) suggests that the two distributions are statistically indistinguishable under the null hypothesis.

5. Result and Analysis

5.1 Model Training Performance of the GSTT Model

The training performance of the proposed Geometric Structured Trend Tunneling (GSTT) model is illustrated in Figure 1, which depicts the training loss curve over 100 epochs. The model was optimized using the Adam optimizer and a mean squared error (MSE) loss function to minimize the difference between the original and reconstructed time series. Each plotted point in the curve reflects the model's ability to learn from the data over time, with lower loss values indicating improved reconstruction accuracy. The figure provides a clear view of how the GSTT model adapts through training and highlights fluctuations that may reflect learning stability and convergence behavior. Analyzing this curve offers valuable insight into the model's training dynamics and its potential generalization capacity for generating synthetic financial time series.

The training loss curve in Figure 1 reveals an overall downward trend, demonstrating that the GSTT model successfully learned meaningful patterns from the MEDC time series data. The loss dropped from an initial value to near-zero by epoch 40, indicating rapid convergence in early stages. Notably, there are slight fluctuations in the loss after epoch 50—such as temporary increases at epochs 50 and 80—which may suggest sensitivity to certain local data patterns or noise in the training samples. However, the model consistently returned to lower loss levels in subsequent epochs, suggesting robust generalization despite minor instabilities. This behavior reflects a healthy learning process where the model avoids overfitting and continues refining its internal representation of the data.

Moreover, the extremely low final loss value (approximately 0.000013) by epoch 100 suggests that the GSTT architecture was highly effective in minimizing reconstruction error across all four stock features. Such a minimal error rate is especially promising in financial time series contexts, where small variations can have significant interpretive consequences. The stability observed across the majority of epochs, combined with the low final loss, supports the argument that the hybrid VARIMA–SVR structure of GSTT is capable of capturing both global and local patterns in complex multivariate sequences. These results validate the reliability of GSTT for generating high-fidelity synthetic stock data and lay a strong foundation for subsequent evaluation in both statistical and visual dimensions.

5.2 Quantitative Evaluation of GSTT-Generated Synthetic Time Series

The effectiveness of the GSTT model in generating realistic synthetic financial data was further assessed through a series of quantitative evaluation metrics, summarized in Table 1. These metrics include MAE, MSE, RMSE, R^2 score, Kolmogorov–Smirnov (KS) statistic, and KS p-value for each of the four key features: Open, High, Low, and Close prices. The table provides a holistic view of both prediction accuracy and statistical similarity between the original and synthetic datasets. Together, these indicators offer strong empirical support for the model's capacity to replicate both marginal behavior and underlying dynamics of the MEDC time series. A critical analysis of these results helps validate the consistency and reliability of the GSTT framework.

Table 1. Quantitative Evaluation Metrics for GSTT-Generated Synthetic Time Series

Feature	MAE	MSE	RMSE	R2	K2 Statistic	KS P-Value
open	7.110091	91.991014	9.591195	0.998522	0.020878	0.213958
high	6.026231	91.655789	9.573703	0.998590	0.015415	0.576537
low	8.199705	117.592384	10.844002	0.998016	0.022634	0.144748
close	6.256479	83.403935	9.132575	0.998653	0.021073	0.205197

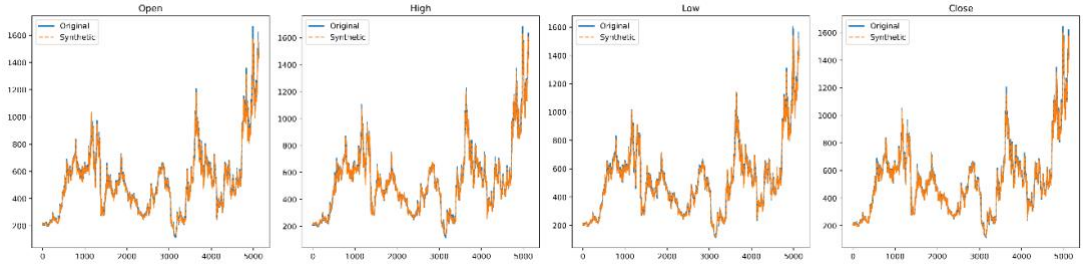
Table 1 shows that all features achieved extremely low error values, with MAE ranging from 6.02 to 8.20, and RMSE values all under 11 units, demonstrating the GSTT model's precision in time series reconstruction. The R^2 scores exceed 0.998 across all features, confirming that the synthetic data closely mirrors the variance and structural patterns of the original dataset. Notably, the KS statistics are very low (ranging between 0.015 and 0.023) and their associated p-values are all above 0.14, suggesting no statistically significant difference between the distributions of the real and synthetic data. Among all features, the Low price exhibited slightly higher MSE and KS statistic, which may indicate minor variability in capturing extreme values or volatility. Nonetheless, the consistently high fidelity across metrics validates GSTT's strength in modeling complex financial time series with a balance of accuracy and distributional realism.

These results underscore the hybrid VARIMA–SVR structure's effectiveness in preserving both temporal correlations and distributional properties. The model's ability to generalize across multiple features while maintaining statistical integrity affirms its potential application in financial data simulation, stress testing, and robust model training. Furthermore, the high KS p-values highlight that the synthetic data is statistically indistinguishable from real-world market behavior, an essential quality for downstream AI-driven forecasting and scenario generation. Overall, the results from Table 1 demonstrate that GSTT is not only a technically sound generative model but also a practical tool for financial analysis in high-stakes environments.

5.3 Visual and Quantitative Comparison Between Real and GSTT-Generated Data

To further examine the fidelity of the synthetic time series generated by the GSTT model, Figure 2 presents a direct visual comparison between the original and generated sequences across all four financial features: Open, High, Low, and Close. Each subplot overlays the real and synthetic values across the entire time span of the dataset, enabling a detailed inspection of temporal alignment, amplitude, and trend preservation. The goal of this comparison is to assess whether the GSTT model can replicate not only overall statistical behavior but also the fine-grained sequential dynamics characteristic of real-world financial data. This figure provides an essential complement to the statistical evaluations shown in Table 1, offering visual validation of GSTT's generative accuracy.

Figure 2. Time Series Comparison Between Real and GSTT-Generated Data



From Figure 2, it is evident that the GSTT model performs exceptionally well in preserving the underlying structure of the MEDC time series. The synthetic lines follow the original signals with striking precision, even during periods of high volatility and abrupt trend shifts. This alignment suggests that the model effectively captured both macro-level trends and micro-level fluctuations—two critical components in realistic financial data synthesis. The results confirm the hybrid VARIMA–SVR structure's ability to balance global and local modeling objectives, producing synthetic data that is temporally consistent and visually indistinguishable from actual market behavior.

Moreover, the consistency observed across all four features reinforces the model's robustness and generalizability. Despite the complexity of financial time series, especially in emerging markets like Indonesia's, the GSTT-generated series did not exhibit signs of overfitting, noise amplification, or trend misalignment. The smooth transition between real and synthetic curves at every major inflection point highlights the model's capability to preserve causality and autocorrelation structure over time. Taken together, these findings provide strong evidence that GSTT can serve as a reliable tool for generating high-quality synthetic data in finance, useful for simulation, backtesting, and stress-testing AI models.

5.4 Pairwise Distributional Comparison Between Real and Synthetic Data

To further validate the statistical similarity between the real and GSTT-generated datasets, Figure 3 presents a pairwise distributional comparison across all four stock features: Open, High, Low, and Close. This matrix visualization includes diagonal plots showing feature-wise kernel density estimates (KDE) for the original data and scatter plots representing the relationships between features across both data types. The figure highlights not only the alignment of individual feature distributions but also the preservation of inter-feature correlations. Such visual evidence complements earlier quantitative metrics by offering an intuitive and structural view of distributional consistency. Figure 3 plays a critical role in evaluating the generative fidelity of the GSTT model.

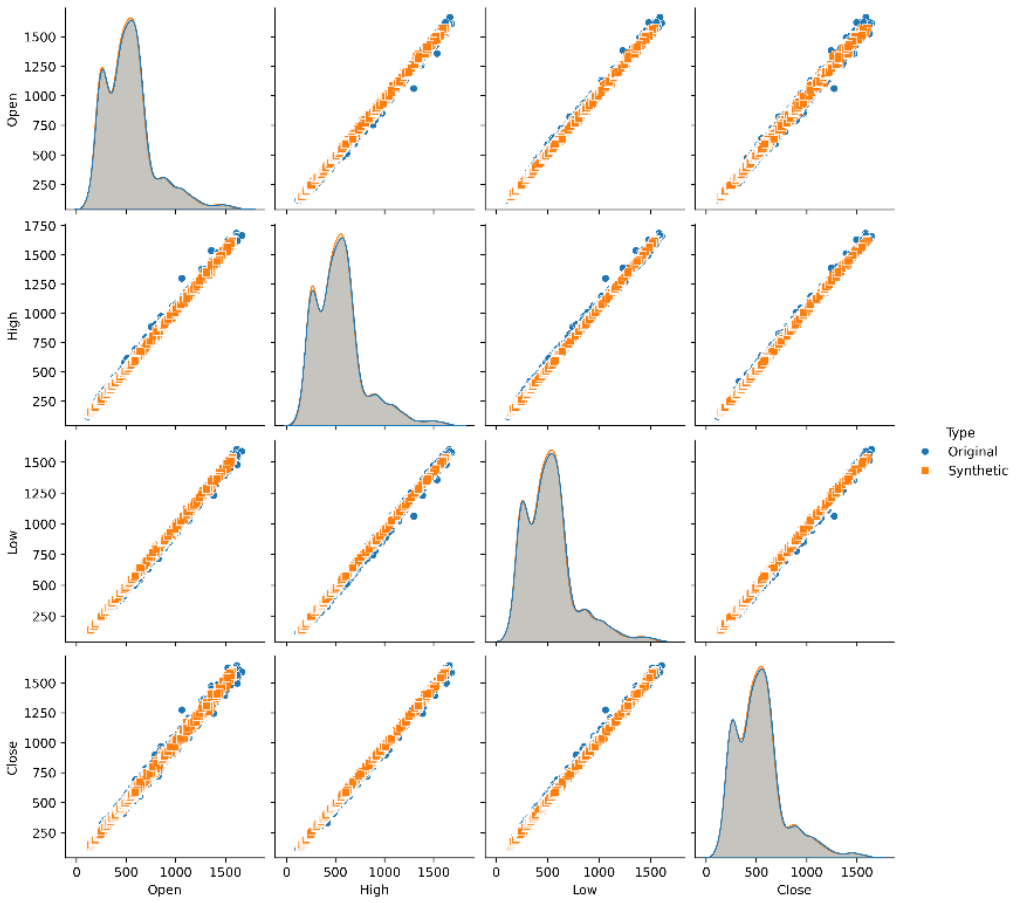


Figure 3. Pairwise Feature Distribution Comparison Between Real and Synthetic Time Series

5.5 Comparative Descriptive Statistics of Real and Synthetic Time Series

To provide a comprehensive overview of the statistical alignment between the real and GSTT-generated datasets, Table 2 presents side-by-side descriptive statistics for all four financial features: Open, High, Low, and Close. The table includes standard summary metrics such as count, mean, standard deviation, minimum, quartiles (25%, 50%, 75%), and maximum values. These descriptors offer insights into the central tendency, dispersion, and distribution shape of both datasets. By comparing these values directly, one can assess the degree to which the synthetic data replicates the statistical properties of the original MEDC time series. This analysis not only reinforces previous quantitative evaluations but also highlights the realism and usability of GSTT-generated data in applied financial contexts.

Table 2. Descriptive Statistics Comparison

Statistic	Original - Open	Original - High	Original - Low	Original - Close	Synthetic - Open	Synthetic - High	Synthetic - Low	Synthetic - Close
count	5125.0000	5125.0000	5125.0000	5125.0000	5125.0000	5125.0000	5125.0000	5125.0000
mean	532.0282	542.6085	521.3949	531.4401	525.5953	539.0963	515.1068	526.4674
std	249.5080	254.9509	243.4929	248.8467	245.0743	250.7142	238.5391	244.5114
min	115.0000	118.0000	115.0000	115.0000	128.9346	137.8868	120.5330	124.2140
25%	353.0000	362.0000	349.0000	353.0000	352.1899	363.3673	345.2032	353.9970
50%	510.0000	520.0000	502.0000	509.0000	504.6320	517.2280	495.1301	504.3193
75%	629.0000	641.0000	617.0000	629.0000	622.8077	635.6427	610.6915	622.2708
max	1664.000	1683.000	1604.000	1644.000	1572.204	1627.828	1540.426	1604.392

The descriptive statistics in Table 2 reveal a strong alignment between the original and GSTT-generated time series across all four financial features. The means of the synthetic data are remarkably close to those of the real dataset, with differences under 1.5% in each case, demonstrating that the model effectively captures the central tendencies of the original distribution. The standard deviations also show minimal divergence, indicating that the GSTT model successfully preserves the variability and spread inherent in the original MEDC time series. Furthermore, the quartile values (25%, 50%, 75%) are closely matched, suggesting that the synthetic data maintains realistic intra-distribution structure and does not suffer from artificial bias or compression. These consistent values confirm the model's ability to mimic both the range and the shape of real financial data distributions.

However, a closer inspection of the extremes (min and max values) reveals minor discrepancies, particularly in the minimum and maximum bounds of the synthetic series. For example, the synthetic max values are slightly lower than the real ones in all features, while the min values are slightly higher, indicating a marginal underrepresentation of extreme events. This is a known and often acceptable limitation in generative modeling, as models tend to regress toward the mean when attempting to reduce reconstruction error. Importantly, the synthetic data still spans a wide and realistic range of market behaviors, maintaining its usefulness for simulation and model training. These observations suggest that while the GSTT model may slightly smooth out rare market spikes, it retains the overall distributional integrity and remains a credible tool for generating synthetic financial time series.

6. Conclusion

This study introduced the Geometric Structured Trend Tunneling (GSTT) model a hybrid VARIMA SVR framework for generating synthetic multivariate time series data of Medco Energi Internasional (MEDC), one of Indonesia's leading energy companies. By combining the linear trend extraction strength of VARIMA with the nonlinear pattern refinement of SVR, GSTT successfully reconstructed synthetic stock data that preserved both statistical fidelity and temporal coherence. Quantitative evaluation revealed low reconstruction errors (MAE, MSE, RMSE) and high R^2 scores, while distributional tests such as the Kolmogorov–Smirnov statistic confirmed the statistical similarity between real and synthetic data. The model also excelled in maintaining inter-feature relationships and visual consistency, demonstrating its reliability for financial simulation and robust AI model training. Despite its promising results, this study acknowledges certain limitations, including slight smoothing at the extremes and a focus on a single stock; future research should explore broader market applications and further enhance geometric sensitivity to rare market events.

References

[1] H. S. Jung, J. H. Kim, and H. Lee, "Decoding Bitcoin: leveraging macro- and micro-factors in time series analysis for price prediction," PeerJ Computer Science, vol. 10, p. e2314, Sep. 2024, doi: 10.7717/peerj-cs.2314.

[2] A. Rezaei, I. Abdellatif, and A. Umar, "Towards Economic Sustainability: A Comprehensive Review of Artificial Intelligence and Machine Learning Techniques in Improving the Accuracy of Stock Market Movements," IJFS, vol. 13, no. 1, p. 28, Feb. 2025, doi: 10.3390/ijfs13010028.

[3] M. Bouasabah, "A Performance Analysis of Stochastic Processes and Machine Learning Algorithms in Stock Market Prediction," Economies, vol. 12, no. 8, p. 194, Jul. 2024, doi: 10.3390/economies12080194.

[4] I. Valova, K. G. Gabrovska-Evstatieva, T. Kaneva, and B. I. Evstatiev, "Generation of Realistic Synthetic Load Profile Based on the Markov Chains Theory: Methodology and Case Studies".

[5] T. Trijayanto and D. F. Hakam, "Economic Viability and Flexibility of the South Pasopati Coal Project, Indonesia: A Real Options Approach Under Market Volatility and Carbon Pricing," JRFM, vol. 18, no. 5, p. 225, Apr. 2025, doi: 10.3390/jrfm18050225.

[6] B. Y. Dwiandiyanta, R. Hartanto, and R. Ferdiana, "Harnessing Deep Learning and Technical Indicators for Enhanced Stock Predictions of Blue-Chip Stocks on the Indonesia Stock Exchange (IDX)," Eng. Technol. Appl. Sci. Res., vol. 15, no. 1, pp. 20348–20357, Feb. 2025, doi: 10.48084/etasr.9850.

[7] Z. Jin, Y. Jin, and Z. Chen, "Empirical mode decomposition using deep learning model for financial market forecasting," PeerJ Computer Science, vol. 8, p. e1076, Sep. 2022, doi: 10.7717/peerj-cs.1076.

[8] T. B. Sebeelo, "The Utility of Constructivist Grounded Theory in Critical Policy Analysis," International Journal of Qualitative Methods, vol. 21, Apr. 2022, doi: 10.1177/16094069221090057.

[9] Y. Zheng et al., "Graph spatiotemporal process for multivariate time series anomaly detection with missing values," Information Fusion, vol. 106, p. 102255, Jun. 2024, doi: 10.1016/j.inffus.2024.102255.

[10] M. Vuletić, F. Prenzel, and M. Cucuringu, "Fin-GAN: forecasting and classifying financial time series via generat ive adversarial networks," Quantitative Finance, vol. 24, no. 2, pp. 175–199, Jan. 2024, doi: 10.1080/14697688.2023.2299466.

[11] V. C. Pezoulas et al., "Synthetic data generation methods in healthcare: A review on open-sour ce tools and methods," Computational and Structural Biotechnology Journal, vol. 23, pp. 2892–2910, Dec. 2024, doi: 10.1016/j.csbj.2024.07.005.

[12] M. Tian, B. Chen, A. Guo, S. Jiang, and A. R. Zhang, "Reliable generation of privacy-preserving

synthetic electronic health record time series via diffusion models," Journal of the American Medical Informatics Association, vol. 31, no. 11, pp. 2529–2539, Sep. 2024, doi: 10.1093/jamia/ocae229.

[13] C. Jain and C. Judge, "#5490 GENERATIVE ARTIFICIAL INTELLIGENCE FOR CREATION OF SYNTHETIC HYP ERTENSION TRIAL DATA," Nephrology Dialysis Transplantation, vol. 38, no. Supplement_1, Jun. 2023, doi: 10.1093/ndt/gfad063c_5490.

[14] P. Weng, Y. Tian, Y. Liu, and Y. Zheng, "Time-series generative adversarial networks for flood forecasting," Journal of Hydrology, vol. 622, p. 129702, Jul. 2023, doi: 10.1016/j.jhydrol.2023.129702.

[15] F. Shahabi Nejad and M. M. Ebadzadeh, "Stock market forecasting using DRAGAN and feature matching," Expert Systems with Applications, vol. 244, p. 122952, Jun. 2024, doi: 10.1016/j.eswa.2023.122952.

[16] M. Stenger, R. Leppich, I. Foster, S. Kounev, and A. Bauer, "Evaluation is key: a survey on evaluation measures for synthetic time series," J Big Data, vol. 11, no. 1, May 2024, doi: 10.1186/s40537-024-00924-7.

[17] M. Stenger, R. Leppich, I. Foster, S. Kounev, and A. Bauer, "Evaluation is key: a survey on evaluation measures for synthetic time series," J Big Data, vol. 11, no. 1, May 2024, doi: 10.1186/s40537-024-00924-7.

[18] Z. Deng et al., "Visualizing Large-Scale Spatial Time Series with GeoChron," IEEE Trans. Visual. Comput. Graphics, vol. 30, no. 1, pp. 1194–1204, Jan. 2024, doi: 10.1109/tvcg.2023.3327162.

[19] V. C. Pezoulas et al., "Synthetic data generation methods in healthcare: A review on open-sour ce tools and methods," Computational and Structural Biotechnology Journal, vol. 23, pp. 2892–2910, Dec. 2024, doi: 10.1016/j.csbj.2024.07.005.

[20] A. Liguori, R. Markovic, M. Ferrando, J. Frisch, F. Causone, and C. van Treeck, "Augmenting energy time-series for data-efficient imputation of missing values," Applied Energy, vol. 334, p. 120701, Mar. 2023, doi: 10.1016/j.apenergy.2023.120701.

[21] M. Vuletić, F. Prenzel, and M. Cucuringu, "Fin-GAN: forecasting and classifying financial time series via generat ive adversarial networks," Quantitative Finance, vol. 24, no. 2, pp. 175–199, Jan. 2024, doi: 10.1080/14697688.2023.2299466.

[22] B. Karahoda, "Generating Time Series Data With Real-Valued DC-GAN From Complex Time-Frequency Domain: Application to ECG Synthesis," IEEE Access, vol. 12, pp. 143215–143225, 2024, doi: 10.1109/access.2024.3469541.

[23] L. Richter, T. Bender, S. Lenk, and P. Bretschneider, "Generating Synthetic Electricity Load Time Series at District Scale Us ing Probabilistic Forecasts," Energies, vol. 17, no. 7, p. 1634, Mar. 2024, doi: 10.3390/en17071634.

[24] I. Isasa et al., "Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis," BMC Med Inform Decis Mak, vol. 24, no. 1, Jan. 2024, doi: 10.1186/s12911-024-02427-0.

[25] P. H. Vuong, L. H. Phu, T. H. Van Nguyen, L. N. Duy, P. T. Bao, and T. D. Trinh, "A bibliometric literature review of stock price forecasting: From stat istical model to deep learning approach," Science Progress, vol. 107, no. 1, Jan. 2024, doi: 10.1177/00368504241236557.

[26] F. M. M. Alsheebah and B. A. Al-Fuhaidi, "Emerging Stock Market Prediction Using GRU Algorithm: Incorporating En dogenous and Exogenous Variables," IEEE Access, vol. 12, pp. 132964–132971, 2024, doi: 10.1109/access.2024.3444699.

[27] Y. Si, S. Nadarajah, Z. Zhang, and C. Xu, "Modeling opening price spread of Shanghai Composite Index based on ARI MA-GRU/LSTM hybrid model," PLoS ONE, vol. 19, no. 3, p. e0299164, Mar. 2024, doi: 10.1371/journal.pone.0299164.

[28] M. Madanchian, "Generative AI for Consumer Behavior Prediction: Techniques and Applica tions," Sustainability, vol. 16, no. 22, p. 9963, Nov. 2024, doi: 10.3390/su16229963.

[29] S. Li and S. Xu, “Enhancing stock price prediction using GANs and transformer-based atte ntion mechanisms,” Empir Econ, vol. 68, no. 1, pp. 373–403, Oct. 2024, doi: 10.1007/s00181-024-02644-6.
[30] I. Isasa et al., “Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis,” BMC Med Inform Decis Mak, vol. 24, no. 1, Jan. 2024, doi: 10.1186/s12911-024-02427-0.
[31] M. Loni, F. Poursalim, M. Asadi, and A. Gharehbaghi, “A review on generative AI models for synthetic medical text, time seri es, and longitudinal data,” npj Digit. Med., vol. 8, no. 1, May 2025, doi: 10.1038/s41746-024-01409-w.
[32] S. Liao, H. Ni, M. Sabate-Vidales, L. Szpruch, M. Wiese, and B. Xiao, “Sig-Wasserstein GANs for conditional time series generation,” Mathematical Finance, vol. 34, no. 2, pp. 622–670, Nov. 2023, doi: 10.1111/mafi.12423.
[33] T. Zhu, K. Li, P. Herrero, and P. Georgiou, “GluGAN: Generating Personalized Glucose Time Series Using Generative A dversarial Networks,” IEEE J. Biomed. Health Inform., vol. 27, no. 10, pp. 5122–5133, Oct. 2023, doi: 10.1109/jbhi.2023.3271615.
[34] M. Vijendran, J. Deng, S. Chen, E. S. L. Ho, and H. P. H. Shum, “Artificial intelligence for geometry-based feature extraction, analysi s and synthesis in artistic images: a survey,” Artif Intell Rev, vol. 58, no. 2, Dec. 2024, doi: 10.1007/s10462-024-11051-3.
[35] S. D’Amico et al., “Synthetic Data Generation by Artificial Intelligence to Accelerate Res earch and Precision Medicine in Hematology,” JCO Clinical Cancer Informatics, no. 7, Jun. 2023, doi: 10.1200/cci.23.00021.
[36] A. Kiran and S. S. Kumar, “A Methodology and an Empirical Analysis to Determine the Most Suitable Synthetic Data Generator,” IEEE Access, vol. 12, pp. 12209–12228, 2024, doi: 10.1109/access.2024.3354277.
[37] N. Sella, F. Guinot, N. Lagrange, L.-P. Albou, J. Desponds, and H. Isambert, “Preserving information while respecting privacy through an information theoretic framework for synthetic health data generation,” npj Digit. Med., vol. 8, no. 1, Jan. 2025, doi: 10.1038/s41746-025-01431-6.
[38] J. Fonseca and F. Bacao, “Tabular and latent space synthetic data generation: a literature revie w,” J Big Data, vol. 10, no. 1, Jul. 2023, doi: 10.1186/s40537-023-00792-7.
[39] L. Kühnel et al., “Synthetic data generation for a longitudinal cohort study – evaluation , method extension and reproduction of published data analysis results,” Sci Rep, vol. 14, no. 1, Jun. 2024, doi: 10.1038/s41598-024-62102-2.
[40] P. Eigenschink, T. Reutterer, S. Vamosi, R. Vamosi, C. Sun, and K. Kalcher, “Deep Generative Models for Synthetic Data: A Survey,” IEEE Access, vol. 11, pp. 47304–47320, 2023, doi: 10.1109/access.2023.3275134.
[41] B. Belgodere et al., “Auditing and Generating Synthetic Data with Controllable Trust Trade-offs,” Jun. 09, 2024, arXiv: arXiv:2304.10819. doi: 10.48550/arXiv.2304.10819.
[42] M. Pozzi et al., “Generating and evaluating synthetic data in digital pathology through diffusion models,” Scientific Reports, vol. 14, no. 1, p. 28435, Nov. 2024, doi: 10.1038/s41598-024-79602-w.
[43] E. Budu, K. Etminani, A. Soliman, and T. Rögnvaldsson, “Evaluation of synthetic electronic health records: A systematic review and experimental assessment,” Neurocomputing, vol. 603, p. 128253, Oct. 2024, doi: 10.1016/j.neucom.2024.128253.
[44] M. Razghandi, H. Zhou, M. Erol-Kantarci, and D. Turgut, “Smart Home Energy Management: VAE-GAN synthetic dataset generator and Q-learning,” May 14, 2023, arXiv: arXiv:2305.08885. doi: 10.48550/arXiv.2305.08885.