Interpretable Deep Learning: Shapley-based Analysis of Generative Models for Synthetic Data Generation

---

A Special Project Presented to the

Faculty of the Department of Computer Science,

College of Science,

University of the Philippines Cebu

In Partial Fulfillment

Of the Requirements for the Degree

Bachelor of Science in Computer Science

Pinky Grace Arcenal Marfa

Bachelor of Science in Computer Science

Asst. Prof. Dharyll Prince M. Abellana

Special Problem Adviser

June 2025

UNIVERSITY OF THE PHILIPPINES CEBU

Bachelor of Science in Computer Science

Pinky Grace Arcenal Marfa

Interpretable Deep Learning: Shapley-based Analysis on Generative Models for Synthetic Data Generation

Asst. Prof. DHARYLL PRINCE M. ABELLANA

Special Problem Adviser

College of Science

Department of Computer Science

Permission is given for the following people to have access to this SP:

Available to the general public	No
Available only after consultation with author/SP adviser	Yes
Available only to those bound by confidentiality agreement	Yes

PINKY GRACE ARCENAL MARFA

Student

Asst. Prof. DHARYLL PRINCE M. ABELLANA

Special Problem Adviser

Approval Sheet

The faculty of the University of the Philippines Cebu – Department of Computer Science approves this Special Problem entitled:

Interpretable Deep Learning: Shapley-based Analysis on Generative Models for Synthetic Data Generation

---

by

PINKY GRACE ARCENAL MARFA

Asst. Prof. DHARYLL PRINCE M. ABELLANA Special Problem Adviser	Date Signed
Asst. Prof. DHARYLL PRINCE M. ABELLANA Chair, Department of Computer Science	Date Signed
Dr. ALVIN G. ROXAS Dean, College of Science	Date Signed

ACKNOWLEDGEMENTS

This study would not have been possible without the guidance, support, and inspiration I received from so many individuals throughout this journey.

First and foremost, I would like to express my deepest gratitude to my SP adviser, Asst. Prof. Dharyll Prince Abellana, whose clarity of thought and brilliance consistently inspired me. His insight, patience, and unwavering belief in my work have shaped this study into what it is today.

To my family, thank you for your never-ending support, encouragement, and love. Your presence has been my foundation through every challenge I faced.

To my friends, acquaintances, and even strangers who, in one way or another, inspired me during my college journey—thank you. Whether through kind words, shared struggles, or silent examples, you've each left a mark that pushed me forward.

And lastly, to myself—for choosing an unfamiliar topic in pursuit of growth, for embracing uncertainty, and for continuing even in moments of doubt. This work is as much a product of my perseverance as it is of academic pursuit.

DEDICATION

To myself, who showed up even in doubt,

and to everyone that follows—

may you find light in the questions,

and courage in the search.

Abstract

This study introduces a Shapley-based evaluation framework for generative models in synthetic time series data generation, addressing the critical gap in interpretable performance attribution across fidelity and usability dimensions. Three generative architectures, TimeGAN, TimeVAE, and VRNNGAN (hybrid), were systematically evaluated across electricity, exchange, and weather datasets using a comprehensive methodology integrating statistical fidelity measures (KL divergence, Wasserstein distance) and forecasting utility metrics (RMSE, MAE). Bootstrap validation with 15 iterations per model-dataset combination provided statistical robustness, while two-way ANOVA analysis revealed significant model-dataset interactions (all $p < 0.001$), confirming that synthetic data generation effectiveness is fundamentally context-dependent. Tukey HSD post-hoc analysis identified distinct performance groupings, with VRNNGAN consistently forming its own superior statistical group across all metrics. Shapley value analysis enabled fair attribution of individual model contributions, revealing that VRNNGAN achieved consistent superiority across 10 of 12 metric-dataset combinations, demonstrating genuine architectural synergy rather than additive benefits. Notably, geometric similarity preservation showed dataset-specific preferences, with TimeVAE excelling for electricity data, TimeGAN for exchange data, and VRNNGAN for weather data. The emergence of negative Shapley values provided novel insights into architectural limitations, indicating that certain model-dataset combinations actively degrade synthetic data quality. This framework advances synthetic data evaluation methodology by providing interpretable insights into when and why specific generative approaches excel, offering practitioners evidence-based model selection guidelines for context-aware synthetic data deployment.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Rationale

The increasing reliance on machine learning (ML) models across various domains such as healthcare, finance, and cybersecurity has fueled an increasing demand for high-quality data that can be used to train and test these systems (Goyal & Mahmoud, 2024; Jordon et al., 2022). However, access to real-world data often comes with significant challenges, including privacy concerns, data scarcity, and algorithmic biases (Hudovernik et al., 2024; Lu et al., 2024). In recent years, synthetic data generation has emerged as a viable solution to these challenges (Goyal & Mahmoud, 2024).

While synthetic data provides a promising alternative to real-world data, its effectiveness largely relies on its quality, which can be assessed based on fidelity and usability. Fidelity evaluates how closely the synthetic data replicates the statistical properties of the original data, while usability assesses its efficacy in supporting ML tasks (Loni et al., 2025). These dimensions are interdependent: high fidelity ideally enhances usability, assuming the original data is reliable. However, if the original dataset contains flaws or biases, replicating it with high fidelity might only perpetuate its limitations (Shahul Hameed et al., 2024). Thus, it is essential for synthetic data to not simply mimic the original data but also address and improve upon its deficiencies to ensure greater usability.

This study aims to evaluate and compare the effectiveness of various generative models in producing synthetic datasets that are both statistically similar to real data (high fidelity) and practically useful in ML applications (high usability). The ultimate goal is to identify models that can best mitigate the shortcomings of low-quality datasets, thereby enhancing the overall robustness and reliability of machine learning systems. Utilizing a Shapley-based analysis, this research quantitatively assesses each generative model's contribution to the quality of synthetic data. This comprehensive evaluation seeks to determine each model's capability not only to adhere closely to the statistical properties of the original data but also to enhance the effectiveness of downstream ML applications. By dissecting the contributions of Generative Adversarial Networks

(GANs), Variational Autoencoders (VAEs), and their combined applications in time series data generation, the study aims to pinpoint which models provide the optimal balance between fidelity and usability.

1.2 Statement of the Problem

The increasing demand for synthetic data in machine learning has led to widespread adoption of generative models such as GANs, VAEs, and hybrid variants. While these models show promise in addressing data scarcity and privacy concerns, evaluating the quality of the synthetic data they generate—particularly for time series—remains a methodological challenge. Most existing studies assess either fidelity or usability in isolation, without offering a unified or interpretable framework that considers both.

Current evaluation practices often rely on black-box comparisons and aggregated metrics, which obscure the contributions of individual model components and mask trade-offs between statistical accuracy and task performance. Consequently, it becomes difficult for practitioners to identify which models work best under specific conditions, or to understand why certain models outperform others. This lack of interpretability limits transparency and trust in generative models, especially in sensitive domains requiring explainable and accountable AI.

The absence of interpretable evaluation methodologies represents a critical gap in synthetic data research. Without systematic frameworks that fairly attribute model contributions across fidelity and usability, practitioners cannot discern context-specific performance patterns or select optimal architectures based on task requirements. Furthermore, aggregate rankings fail to explain model performance, hindering evidence-based decisions and raising concerns about the reliability of synthetic data, particularly in domains that demand trust and accountability.

1.2.1 Research Questions

1. How can Shapley value analysis provide interpretable insights into the fidelity and usability contributions of generative models?

2. How do model-dataset interactions affect the interpretability of fidelity-usability trade-offs?

3. Which generative model architecture provides optimal performance across varying conditions?

1.3 Significance of the Study

This study addresses a critical gap in synthetic data generation methodology by introducing the first Shapley-based evaluation framework specifically designed for generative models in time series applications. While existing research has focused on isolated performance comparisons between GANs, VAEs, and hybrid architectures, there has been no systematic approach that provides interpretable attribution of each model's contribution to synthetic data quality across both fidelity and usability dimensions.

The Shapley-based framework represents a novel application of cooperative game theory principles to generative model assessment, enabling fair attribution of performance contributions rather than traditional black-box comparisons. This systematic integration of fidelity and usability metrics within a unified evaluation framework addresses the longstanding challenge of balancing statistical accuracy with practical utility in synthetic data applications. This framework enables evidence-based model selection guidelines that move beyond aggregate performance metrics to offer nuanced understanding of when and why specific generative approaches excel. The methodology supports identification of optimal models based on target application requirements and dataset characteristics, facilitating more informed architectural decisions in generative model development. Additionally, the balanced evaluation framework offers actionable insights for quality assessment that consider both statistical fidelity and downstream task performance, enabling confident deployment of synthetic data in real-world applications.

Ultimately, this research advances the field of synthetic data generation by reframing evaluation as an interpretable, context-aware process. It challenges the assumption of universal model performance and calls for dataset-sensitive assessment strategies that align with the growing demand for trustworthy and explainable AI systems.

1.4 Scopes and Delimitations of the Study

This study focuses on evaluating the quality of synthetic data generated by specific generative models—Generative Adversarial Networks (GANs), Variational Autoencoders

(VAEs), and their hybrid (GAN-VAES)—within the context of time series applications. The primary evaluation framework integrates the concept of fidelity and usability. These two concepts are treated as interrelated aspects of data quality, rather than as separate, isolated evaluative criteria.

1.4.1 Delimitations

- **Privacy Metrics:** While privacy is crucial for synthetic data, especially in regulated fields, this study excludes privacy metrics to maintain a clear focus on evaluating synthetic data quality through fidelity and usability. Integrating privacy metrics would broaden the scope of the research and require more complex methodologies, such as differential privacy techniques, which would complicate the evaluation process and possibly overshadow the evaluation of the other metrics. The exclusion of privacy metrics allows this research to provide a more concentrated and detailed analysis of how well synthetic data replicates statistical properties and supports machine learning tasks. By focusing exclusively on fidelity and usability, the study aims to offer deeper insights into the core aspects of synthetic data quality without the confounding effects of integrating privacy considerations, which could be addressed in future research.

- **Generative Models:** The study limits its analysis to GANs, VAEs, and their hybrid GAN-VAEs. Other types of generative models, such as Restricted Boltzmann Machines or Diffusion Models, are not included in this assessment. This delimitation is based on the established capability of the selected models to handle time series data effectively, which is central to the research questions posed.

- **Dataset Size:** To manage computational costs and training time, the study employs datasets standardized to contain a maximum of 5000 samples. This ensures consistent and efficient evaluation across all models without compromising the integrity of the results.

- **General Applications:** The study focuses on general synthetic data generation for time series applications and is not tailored to specific domains such as healthcare or finance. While the methods used in this research could be adapted for domain-specific applications, the findings are intended to provide insights that are

broadly applicable.

- **Evaluation Focus:** The study does not aim to separately optimize fidelity and usability metrics but rather examines the balance and trade-offs between them in producing high-quality synthetic data. This approach is chosen to better understand how improvements in one aspect might affect the other, providing a holistic view of model performance.

1.4.2 Limitations

- **Dataset Representation:** The findings are limited to the datasets used in this study, which include time series datasets of varying dimensionality. These datasets, selected for their relevance to existing literature and practicality, may not fully represent the diversity and complexity of real-world time series data.

- **Evaluation Scope:** The evaluation is limited to forecasting tasks using LSTM as the downstream application for utility assessment, which may not generalize to other machine learning tasks such as classification or clustering.

- **Computational Constraints:** The decision to limit dataset size and use simpler forecasting models reflect practical constraints and may not capture the full potential of generative models under larger-scale experiments.

- **Exclusion of Privacy Metrics:** While privacy metrics are not included in this study, their absence represents a potential area for future exploration. This limitation reflects a deliberate decision to focus on fidelity and usability, which are critical dimensions of synthetic data quality. Privacy-preserving synthetic data generation is only meaningful if the synthetic data itself has high fidelity and usability. By ensuring that the foundation of synthetic data quality is addressed first, this study creates a strong basis for exploring privacy considerations in future work. However, the exclusion of privacy metrics means this study does not evaluate the extent to which the generated datasets mitigate risks of privacy breaches, leaving this as an area for further exploration.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of Synthetic Data Generation

Synthetic data refers to artificially generated data produced through specialized mathematical models or algorithms, specifically designed to address particular data science tasks (Jordon et al., 2022). According to Jordon et al. (2022), synthetic data serves as a tool to address various challenges in data science, including privacy concerns, bias, and data scarcity, without directly exposing sensitive information. This view is further supported by Goyal and Mahmoud (2024) who have highlighted the increasing recognition of synthetic data for its potential to address pressing real-world issues such as mitigating data scarcity, addressing privacy concerns, and reducing algorithmic biases— issues common in machine learning applications. Similarly, Emam et al. (2020) emphasizes that although it is not real data, synthetic data is generated based on the statistical properties of the original dataset, ensuring it mirrors the original data in terms of patterns and distributions. This approach enables for the preservation of statistical integrity while mitigating risks related to data privacy breaches.

Expanding on these definitions, synthetic data has seen widespread adoption in a variety of fields, where the aforementioned challenges are most pronounced. This is especially evident in the healthcare domain, where access to relevant datasets is often constrained due to privacy and scarcity. A study by Skandarani et al. (2021) illustrates this through the use of Generative Adversarial Networks (GANs) to generate medical images to enable meaningful research. This application is further elaborated by Habiba et al. (2021), who investigated ECG synthesis using Neural Ordinary Differential Equations (ODE) and GAN models, demonstrating another facet of how synthetic data can support advancements in medical research. Similarly, in other domains like finance, where there is a need for balanced datasets and anonymization, synthetic data has proven beneficial as exemplified by Caliskan et al. (2023), comparatively analyzing the Variational Auto-Encoder (VAE) and Conditional Tabular Generative Adversarial Network (CTGAN) for generating synthetic financial credit load data. Additionally, synthetic data is also used in network traffic simulation (Cullen et al., 2022), supporting

cybersecurity applications through the generation of anonymized datasets.

Despite its utility, there remains significant open challenges in the field of synthetic data generation. A recurring theme across the related literature is the lack of well-agreed metrics for evaluating the quality, utility, and privacy of synthetic data (Bauer et al., 2024; Caliskan et al., 2023; Goyal & Mahmoud, 2024; Lu et al., 2024). This absence of well-agreed evaluation methods makes it a challenging task to compare the performance of different models, such as GANs, VAEs, and Neural ODEs, and to determine which model is best suited for a given application, as observed in a systematic review by Goyal and Mahmoud (2024). For example, while some studies rely on statistical difference evaluations like Kullback-Leibler (KL) divergence or Wasserstein distance (Li et al., 2019), others employ human evaluation or application-specific metrics to assess synthetic data quality (Lu et al., 2024), which further contributes to the inconsistencies in evaluation. Beyond evaluation, the balance between fidelity and privacy also remains a critical issue, with multiple studies emphasizing the difficulty of generating data that mirrors the original dataset while protecting sensitive information (Goyal & Mahmoud, 2024). Additional issues include propagation of biases (Jordon et al., 2022), fairness issues (Lu et al., 2024), high computational costs (Bauer et al., 2024) and the lack of robust privacy-preserving techniques (Kaabachi et al., 2024).

In light of the challenges discussed above, it is evident that the field of synthetic data generation is at a crucial juncture. Given the growing reliance on machine learning models across sectors, understanding how to measure the quality and impact of synthetic data will become increasingly important. Thus, the study of metrics for evaluating synthetic data—in terms of both privacy protection and data utility—will be a critical area of future research. Such metrics will not only improve synthetic data's applicability across industries but also enhance trust in its use for privacy-sensitive applications like healthcare and finance.

2.2 Metrics for Evaluating Synthetic Data

As the field of synthetic data generation evolves, the need for robust and widely-accepted metrics to evaluate the quality, utility, and privacy of synthetic data has become increasingly apparent. These metrics are essential not only for ensuring that synthetic data serves practical purposes but also for guaranteeing its ethical use in sensitive appli-

cations such as healthcare (Kaabachi et al., 2024). Evaluation metrics are crucial for determining how well synthetic data replicates the original dataset while maintaining privacy, ensuring the data's usability, and avoiding unintended consequences such as the leakage of sensitive information (Jordon et al., 2022; Kaabachi et al., 2024). In the literature, synthetic data evaluation metrics are broadly categorized into two categories: utility metrics and privacy metrics (Goncalves et al., 2020; Kaabachi et al., 2024).

Utility metrics play a key role in assessing how well synthetic data retains the essential characteristics of the original data, ensuring its usefulness for tasks such as machine learning model training and data analysis. As defined by Hittmeir et al. (2019), utility metrics quantify how useful synthetic data is by assessing how much information is lost during the data generation process. In this way, they help determine how closely the synthetic data mirrors the original, supporting its use in real-world applications. These utility metrics can be further divided into general and task-specific metrics (Kaabachi et al., 2024; Osorio-Marulanda et al., 2024).

General utility metrics assess the overall statistical properties and model evaluation results for a wide range of potential analyses that could be performed on the data (El Emam, 2020). These metrics focus on preserving key statistical attributes of the original dataset, such as the distribution of variables, means, variances, and correlations. Commonly used general utility metrics include KL Divergence and Wasserstein Distance, both of which measure the similarity between the probability distributions of the real and synthetic data (Fonseca & Bacao, 2023). While general utility metrics provide a broad assessment of how well synthetic data mirrors the statistical properties of the original dataset, task-specific utility metrics evaluate synthetic data based on how well it supports specific tasks, such as machine learning applications. These metrics focus on the performance of models trained on synthetic data when applied to real-world tasks. Common task-specific utility metrics used in research include machine learning efficacy metrics like accuracy, precision, recall, and F1-score (Figueira & Vaz, 2022). This approach is particularly useful in classification tasks, where the goal is to determine whether the synthetic data can be used to train models that perform comparably to those trained on real data (Kaabachi et al., 2024).

In addition to utility metrics, privacy metrics are crucial for evaluating the extent to which synthetic data protects sensitive information in the original dataset. Differential

Privacy (DP) is one of the most widely recognized privacy metrics, providing a formal framework that quantifies the privacy guarantees of a data release mechanism (Goyal & Mahmoud, 2024; Jordon et al., 2022; Kaabachi et al., 2024; Nikolenko, 2019). Similarly, Distance to Closest Record (DCR) offers another valuable perspective by measuring the distance between each synthetic data point and its nearest neighbor in the real training dataset (Mendelevitch & Lesh, 2021).

In evaluating synthetic data generation, one of the most significant challenges is the inconsistency in comparative evaluations of different methods, which arises from the lack of uniform evaluation metrics (Chundawat et al., 2024). This inconsistency leads to confusion and variability in determining the effectiveness of synthetic data. As different models and applications prioritize various aspects, such as privacy, accuracy, or utility, establishing a common ground for evaluation becomes problematic. For example, while some methods may focus on enhancing privacy, others might aim to improve the accuracy or utility of the synthetic data, resulting in a diverse range of metrics that are not easily comparable. In response to these discrepancies, there have been concerted efforts within the research community to develop more unified evaluation metrics. One notable example is TabSynDex by Chundawat et al. (2024), a single-score metric designed to robustly evaluate synthetic tabular data. Another innovative metric that has been developed is SHAPr (Duddu et al., 2022), a Shapley-value based privacy metric which offers a versatile tool for measuring privacy risks.

However, these efforts encounter additional complexities. A prominent issue is the difficulty in balancing the trade-offs between privacy and utility. Enhancing privacy typically involves introducing mechanisms that may degrade the utility of the data by reducing its accuracy or limiting the types of analysis that can be performed effectively. This trade-off is particularly problematic in fields requiring high-fidelity data for accurate analysis, such as healthcare and finance, where both high utility and stringent privacy are critical (Caliskan et al., 2023; Mendelevitch & Lesh, 2021). Moreover, many existing studies and methodologies fail to thoroughly evaluate residual privacy risks, especially concerning publicly released synthetic data (Kaabachi et al., 2024). This oversight can lead to significant privacy breaches, as residual information may still be exploited to uncover sensitive data about individuals in the original dataset. Lastly, the metrics currently employed often struggle to capture complex relationships and depen-

dencies between multiple attributes in the data. This deficiency can lead to synthetic datasets that, while statistically similar to original datasets in marginal distributions, fail to preserve more complex interdependencies, thus limiting their usefulness for more sophisticated data science tasks (Jordon et al., 2022).

Table 1: Key Evaluation Metrics in Synthetic Data Generation Literature

Category	**Metric**	**Application**	**Source**
Fidelity	KL Divergence	Measures distributional similarity between synthetic and original data	Li et al. (2019)
	Wasserstein Distance	Evaluates geometric similarity preservation in data distributions	Li et al. (2019)
Utility	RMSE	Emphasizes larger prediction errors for forecasting evaluation	Hazra et al. (2022)
	MAE	Measures average absolute prediction errors	Hazra et al. (2022)
	Accuracy	Measures proportion of correct predictions	Figueira and Vaz (2022)
	F1-Score	Provides balanced score of precision and recall	Figueira and Vaz (2022)
Privacy	Distance to Closest Record	Evaluates privacy by assessing similarity to nearest original data points	Mendelevitch and Lesh (2021)
	Differential Privacy	Formal privacy measure that quantifies privacy risk	Nikolenko (2019)
Unified	TabSynDex	Single-score metric for tabular data fidelity and utility	Chundawat et al. (2024)
	SHAPr	Shapley-based privacy risk assessment	Duddu et al. (2022)

2.3 Survey of Generative Deep Learning Models

Generative deep learning models have emerged as powerful tools for creating synthetic data, increasingly becoming relevant in fields like healthcare, finance, and security, where

data privacy and scarcity are major concerns. These models aim to capture the underlying data distribution of the training set in order to produce samples that reflect the learned distribution (Carvajal-Patiño & Ramos-Pollán, 2022). For example, GANs use an adversarial framework where the generator learns to model the data distribution by producing synthetic samples that fool a discriminator, which is trained to distinguish real from synthetic data. Similarly, VAEs utilize an encoder-decoder architecture, where the encoder maps data into a latent space and the decoder generates new samples by reconstructing data from this latent representation. Prior to the development of deep learning models, early generative approaches, such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) paved the way by capturing simple probabilistic relationships in data. However, as Cao et al. (2023) notes, it was the advent of deep learning that brought major performance advancements, allowing generative models to handle complex, high-dimensional data distributions. This idea is supported by Caliskan et al. (2023), who suggest that the proliferation of deep learning algorithms and emergence of generative techniques offers more promising solutions in the generation of financial credit loan data. This leap in capability has established deep learning-based models as central to modern synthetic data generation.

With the rise of deep learning, models such as GANs and VAEs have emerged as foundational generative models that leverage neural networks to capture and reproduce complex data distributions in high-dimensional spaces. GANs, introduced by Goodfellow et al. (2014), use an adversarial framework involving two neural networks, a generator and a discriminator. More specifically, the generator begins with random noise as its input and strives to produce data whose distribution challenges the discriminator's ability to classify it as real or synthetic. Based on the discriminator's classification results, the gradients of the generator are updated to better approximate the real data distribution (Zia et al., 2023). Although initially focused on synthetic image generation, GANs have extended their utility beyond image synthesis to domains like time-series data generation (Yoon et al., 2019).

VAEs represent another class of generative models frequently used for synthetic data generation. Proposed by Kingma and Welling (2022), VAEs use a probabilistic approach, learning to encode input data into a latent space in a manner that enables controlled sampling. Unlike GANs, which rely on adversarial training, VAEs employ a

probabilistic approach using encoder and decoder networks to learn and generate data (Lu et al., 2024). One of the key strengths of VAEs lies in their ability to generate smooth, continuous latent spaces. This quality enables fine control over the generative process, allowing the user to manipulate specific features or generate data with particular attributes by sampling specific regions of the latent space, such as in anomaly detection (Niu et al., 2020), where VAEs can learn normal patterns and identify deviations as anomalies.

In addition to GANs and VAEs, Restricted Boltzmann Machines (RBMs) have also been critical in the evolution of generative models. Proposed by Salakhutdinov et al. (2007), RBMs are models that can learn the distribution of the training data through a two-layer architecture — a visible layer that represents the observed data and a hidden layer that captures latent features (Carvajal-Patiño & Ramos-Pollán, 2022). RBMs have been successfully applied in a range of tasks, including generating user preferences for recommendation systems like Netflix (Nematholahy, 2020) and reconstructing missing data in tabular datasets. For instance, in financial modeling, they have been used to generate synthetic market data, replicating the probability distributions of real-world datasets and capturing complex dependencies (Kondratyev & Schwarz, 2019).

Originally introduced by Sohl-Dickstein et al. (2015), Diffusion Models have gained significant attention in synthetic data generation in recent years due to their ability to handle complex data distributions with stability and precision. As per Lin et al. (2023), the underlying principle of Diffusion Models is to progressively perturb observed data through a forward diffusion process and then recover the original data using a backward reverse process. The forward process involves multiple steps of noise injection, where the noise level changes incrementally at each step. Conversely, the backward process consists of a series of denoising steps, parameterized by a neural network, that gradually remove the injected noise. Once the backward process has been learned, Diffusion Models can generate new samples from almost any initial data (Lin et al., 2023). This is supported by a study by You et al. (2023) that highlights that Diffusion Models are particularly effective in scenarios with limited training data, as they can generate novel samples even when very few training samples are available. In addition, Zhu (2024) emphasizes that Diffusion Models, along with other foundational generative models like GANs, have become one of the most widely used methods for modeling the distribution of

continuous-domain data and generating new samples. Their growing popularity reflects their versatility and robustness in a wide range of applications, especially as Diffusion Models have demonstrated their power over many existing generative techniques (Lin et al., 2023).

CHAPTER 3

METHODOLOGY

This study follows a structured methodological process that involves data preparation, evaluation of generative models, and analysis of their contributions to synthetic data quality using Shapley value analysis. The focus is on assessing fidelity and usability metrics to determine which generative model provides the best balance between these dimensions. Each step is further elaborated in the following subsections.

3.1 Dataset Preparation

For this study, three multivariate datasets are considered. These datasets were chosen due to their widespread use in forecasting tasks in the literature, ensuring consistency and providing context for evaluating the generative models under study.

1. **Exchange:**

   This dataset, sourced from the study by Lai et al. (2017), consists of 7,587 daily exchange rate records across eight foreign currencies: Australian Dollar, British Pound, Canadian Dollar, Swiss Franc, Chinese Yuan, Japanese Yen, New Zealand Dollar, and Singapore Dollar, spanning the period from 1990 to 2016. For this study, only the British Pound and Japanese Yen exchange rates were selected using seeded randomization.

2. **Electricity:**

   Obtained from Zhou et al. (2020), this dataset comprises 17,420 samples with 7 features, representing two years of electricity transformer load data from two regions in a Chinese province. All features are used in the model evaluation. This dataset is a widely-used benchmark, cited in 27 benchmark studies and referenced in 288 papers between 2020 and 2025.

3. **Weather:**

   This meteorological dataset contains 52,696 samples with 21 features recorded every 10 minutes throughout 2020. Ten features were selected for this study using seeded randomization. The dataset represents a commonly used benchmark in

time series forecasting research, with documented usage across multiple comparative studies.

To maintain consistency in training and evaluation, all datasets will be adjusted to contain the latest 5000 samples. This ensures comparable results across datasets while reducing computational costs during training. Additionally, all datasets will undergo Min-Max normalization and handling of missing values using forward-fill methods, if any missing values are present.

3.2 Generative Models

This study evaluates three generative model approaches: TimeGAN, TimeVAE, and VRNNGAN (a hybrid architecture). These models were selected for their established effectiveness in time series synthetic data generation and their complementary approaches to capturing temporal dependencies.

1. **TimeGAN**

   TimeGAN (Yoon et al., 2019) represents the adversarial approach to time series generation, utilizing a generator-discriminator framework specifically designed for sequential data. The model incorporates temporal dynamics through recurrent neural networks while maintaining the adversarial training paradigm that enables realistic synthetic sequence generation.

2. **TimeVAE**

   TimeVAE (Desai et al., 2021) employs a variational autoencoder architecture adapted for time series data, learning probabilistic latent representations that capture temporal patterns. The model's encoder-decoder structure with variational inference enables controlled generation of synthetic sequences from learned latent distributions.

3. **VRNNGAN (Hybrid)**

   VRNNGAN (Lee, 2022) combines the strengths of both adversarial and variational approaches, integrating the structured latent space learning of VAEs with the realistic sample generation capabilities of GANs. This hybrid architecture aims to balance statistical fidelity with practical utility for downstream applications.

3.3 Experimental Design and Statistical Validation

3.3.1 Bootstrap Methodology

To ensure statistical robustness and reliable performance estimates, the evaluation phase employed bootstrap sampling of the 30% testing data. Each model-dataset combination was evaluated through 15 bootstrap runs, where different bootstrap samples of the test set were used to assess synthetic data quality, generating a total of 135 observations across all experiments (3 models $\times$ 3 datasets $\times$ 15 runs).

The bootstrap evaluation procedure was implemented as follows:

1. Models were trained once on the 70% training portion of each dataset
2. For each bootstrap iteration, random sampling with replacement was performed on the 30% test set to create bootstrap test samples
3. Each trained generative model generated synthetic data based on the training set
4. Evaluation of synthetic data quality was conducted using the bootstrap test samples for all four metrics (KL divergence, Wasserstein distance, RMSE, MAE)
5. Performance measurements were recorded for statistical analysis across the 15 bootstrap iterations

This bootstrap approach provides robust estimates of model performance while accounting for testing variability and ensuring that results are not dependent on specific test data partitions. By bootstrapping the evaluation phase rather than the training phase, the methodology maintains consistent model training while generating reliable confidence intervals for performance assessment.

3.3.2 Statistical Analysis Framework

The experimental design employs a two-way factorial ANOVA to examine the effects of model type (TimeGAN, TimeVAE, VRNNGAN) and dataset characteristics (Electricity, Exchange, Weather) on synthetic data quality metrics. This analysis tests for:

- Main effects of generative model type
- Main effects of dataset characteristics

- Model × Dataset interaction effects

Following significant ANOVA results, Tukey HSD post-hoc tests are conducted to identify specific pairwise differences between models, providing detailed insights into performance hierarchies across different evaluation contexts.

3.4 Evaluation Metrics

3.4.1 Fidelity Evaluation

The fidelity of synthetic data is assessed through distributional similarity measures that quantify how closely the generated data replicates the statistical properties of the original dataset:

1. **Kullback-Leibler (KL) Divergence:** Measures the divergence between probability distributions of synthetic and real data, quantifying distributional similarity with lower values indicating better fidelity.

2. **Wasserstein Distance:** Evaluates the earth mover's distance between synthetic and real data distributions, providing insights into geometric similarity with lower values indicating better preservation of distributional characteristics.

3.4.2 Usability Evaluation

Usability assessment focuses on the practical utility of synthetic data for downstream forecasting tasks using a Train on Synthetic-Test on Real (TSTR) paradigm (Hernandez et al., 2022):

1. **Forecasting Setup:** Long Short-Term Memory (LSTM) networks are trained exclusively on synthetic data generated by each model and evaluated on held-out real data to assess practical utility.

2. **Performance Metrics:**

   - **Root Mean Square Error (RMSE):** Emphasizes larger prediction errors through squared deviations, providing insights into forecasting robustness.
   - **Mean Absolute Error (MAE):** Quantifies average prediction accuracy through absolute deviations, offering overall measure of forecasting utility.

3.4.3 Model Configuration and Hyperparameters

To ensure reproducibility and transparency, all generative models were implemented using their default configurations as specified in their respective original publications and library implementations. The LSTM forecasting model used for usability evaluation was configured based on established best practices for time series forecasting tasks.

TimeGAN and TimeVAE were implemented using the Synthcity library (Qian et al., 2023), which provides standardized implementations of synthetic data generation models. For TimeGAN, the base RNN architecture was modified to LSTM to ensure compatibility with the forecasting evaluation framework and maintain consistency with the LSTM-based usability assessment. TimeVAE utilized the standard Synthcity configuration with its default variational autoencoder parameters optimized for time series generation. VRNNGAN was implemented using the original author's implementation from Lee (2022), following the hybrid architecture specifications that combine both GAN and VAE components with their respective default settings. Sequence lengths were adapted to each dataset's characteristics for all models to optimize temporal pattern recognition while maintaining comparability across architectures.

The decision to utilize default parameters for all generative models ensures reproducibility and eliminates potential bias that could arise from model-specific hyperparameter optimization. This approach maintains consistency with the original implementations while providing fair comparative evaluation across different architectural paradigms, ensuring that performance differences reflect inherent model capabilities rather than optimization advantages.

The LSTM forecasting model employed a simple shallow architecture with 64 hidden units, following established guidelines for time series forecasting where moderate network complexity often provides optimal balance between learning capacity and overfitting prevention. The choice of 64 hidden units aligns with recommendations by Prihatno et al. (2021), who demonstrated that this configuration provides sufficient representational capacity for time series prediction tasks while avoiding the overfitting and underfitting risks. The shallow LSTM architecture was deliberately chosen to focus evaluation on synthetic data quality rather than forecasting model sophistication, ensuring performance differences reflect data utility rather than model complexity. The complete

LSTM configuration is detailed in Table 2.

Table 2: LSTM Forecasting Model Configuration

Parameter	**Value**
Hidden Units	64
Number of LSTM Layers	1
Fully Connected Layers	1
Sequence Length	Variable*
Batch Size	32
Learning Rate	0.002
Training Epochs	50
Optimizer	Adam
Loss Function	MSE
Prediction Horizon	1 step ahead
Architecture Type	Shallow LSTM

**Sequence length varies by dataset characteristics*

3.5 Shapley Value Analysis Framework

The Shapley value analysis provides fair attribution of each model's contribution to synthetic data quality by considering all possible coalitions and marginal contributions. This approach builds upon the foundational work of Shapley (1952), who established the mathematical framework for fairly distributing gains among players in cooperative games based on their marginal contributions across all possible coalition formations.

3.5.1 Implementation

The Shapley value analysis is implemented separately for each metric-dataset combination to preserve context-dependent performance patterns. Since the metrics used are error-based—where lower values indicate better performance—a transformation is applied to convert these into utility values, where higher is better. This enables compatibility with the additive nature of Shapley values.

The utility for a model i is defined as:

$$u_i = \max_j(p_j) - p_i$$

where p_i is the raw performance (error) of model i, and $\max_j(p_j)$ is the worst (highest) performance observed across all models. This transformation ensures that the worst-performing model receives zero utility, and others receive values proportional to their improvement over the worst case.

The characteristic function $v(S)$ represents the performance utility achieved when using the generative approaches specified in coalition S, defined as:

- $v(\emptyset) = 0$ (empty coalition baseline)
- $v(\{\text{GAN}\})$ = utility contribution of TimeGAN individually
- $v(\{\text{VAE}\})$ = utility contribution of TimeVAE individually
- $v(\{\text{GAN-VAE}\})$ = utility contribution of the hybrid VRNNGAN architecture

For the two-player Shapley framework with TimeGAN and TimeVAE as individual players and VRNNGAN as their collaborative outcome, the Shapley contributions are calculated using closed-form expressions:

$$\phi_{\text{GAN}} = \frac{1}{2}\big[v(\{\text{GAN}\}) - v(\emptyset)\big] + \frac{1}{2}\big[v(\{\text{GAN-VAE}\}) - v(\{\text{VAE}\})\big],$$

$$\phi_{\text{VAE}} = \frac{1}{2}\big[v(\{\text{VAE}\}) - v(\emptyset)\big] + \frac{1}{2}\big[v(\{\text{GAN-VAE}\}) - v(\{\text{GAN}\})\big],$$

$$\phi_{\text{GAN-VAE}} = v(\{\text{GAN-VAE}\}) - v(\emptyset).$$

These formulas ensure fair attribution by averaging each player's marginal contributions across all possible orderings of coalition formation. The first term in each individual player's formula represents their standalone contribution, while the second term captures their marginal contribution when joining the other player. The hybrid model's Shapley value represents its total utility contribution relative to the baseline.

Positive Shapley values indicate performance improvement, while negative values suggest performance degradation. The analysis enables identification of context-dependent model strengths and provides guidance for dataset-specific model selection. The higher the Shapley value, the greater the model contributes to the quality of synthetic data generated.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents and analyzes the comprehensive findings derived from the Shapley-based evaluation framework applied to generative model performance in synthetic time series data generation. The analysis encompasses bootstrap validation results that establish statistical robustness, ANOVA findings that reveal significant model-dataset interactions, and Shapley value attributions that provide interpretable insights into individual model contributions across fidelity and usability dimensions.

4.1 Bootstrap Analysis Overview

To ensure statistical robustness, each model-dataset combination was evaluated through 15 bootstrap runs, generating a total of 135 observations across all experiments. This approach provides sufficient statistical power for robust ANOVA testing and reliable confidence interval estimation.

The bootstrap results (Tables 3, 4, and 5) demonstrate consistent performance patterns across all datasets. VRNNGAN consistently achieves the lowest standard deviations across most metrics, indicating stable performance regardless of data sampling variations. All models show tight 95% confidence intervals, confirming the reliability of the performance estimates used in subsequent statistical analyses.

Notably, the bootstrap analysis reveals dataset-specific performance variations that justify the need for individual dataset evaluation rather than aggregated comparisons.

4.1.1 Electricity Dataset Bootstrap Summary

The electricity dataset shows the most consistent model performance, with all models demonstrating tight confidence intervals and minimal variability across bootstrap runs.

Table 3: Bootstrap Summary Statistics for Electricity Dataset (15 runs per model)

Metric	**Model**	**Mean**	**Std Dev**	**95% CI**	**Range**
KL Divergence	TimeGAN	13.268	0.031	[13.251, 13.285]	[13.218, 13.326]
	TimeVAE	13.620	0.043	[13.597, 13.644]	[13.556, 13.690]
	VRNNGAN	0.883	0.014	[0.876, 0.891]	[0.859, 0.908]
Wasserstein Distance	TimeGAN	0.161	0.001	[0.160, 0.161]	[0.159, 0.163]
	TimeVAE	0.102	0.001	[0.101, 0.102]	[0.099, 0.104]
	VRNNGAN	0.125	0.001	[0.125, 0.126]	[0.123, 0.127]
RMSE	TimeGAN	0.245	0.001	[0.245, 0.246]	[0.243, 0.248]
	TimeVAE	0.206	0.001	[0.205, 0.206]	[0.204, 0.207]
	VRNNGAN	0.186	0.002	[0.186, 0.187]	[0.183, 0.189]
MAE	TimeGAN	0.189	0.001	[0.188, 0.189]	[0.187, 0.191]
	TimeVAE	0.170	0.001	[0.170, 0.171]	[0.169, 0.172]
	VRNNGAN	0.125	0.001	[0.125, 0.126]	[0.122, 0.127]

4.1.2 Exchange Dataset Bootstrap Summary

Exchange data reveals slightly higher variability for VRNNGAN in KL divergence (std: 0.037), while maintaining stable performance across other metrics.

Table 4: Bootstrap Summary Statistics for Exchange Dataset (15 runs per model)

Metric	**Model**	**Mean**	**Std Dev**	**95% CI**	**Range**
KL Divergence	TimeGAN	17.291	0.158	[17.203, 17.378]	[16.869, 17.489]
	TimeVAE	16.534	0.158	[16.446, 16.621]	[16.107, 16.734]
	VRNNGAN	0.964	0.037	[0.944, 0.985]	[0.890, 1.032]
Wasserstein Distance	TimeGAN	0.117	0.003	[0.116, 0.119]	[0.112, 0.121]
	TimeVAE	0.181	0.003	[0.179, 0.182]	[0.174, 0.186]
	VRNNGAN	0.172	0.003	[0.171, 0.174]	[0.166, 0.177]
RMSE	TimeGAN	0.187	0.003	[0.185, 0.188]	[0.181, 0.192]
	TimeVAE	0.240	0.004	[0.238, 0.242]	[0.231, 0.245]
	VRNNGAN	0.171	0.004	[0.169, 0.173]	[0.166, 0.178]
MAE	TimeGAN	0.144	0.002	[0.142, 0.145]	[0.139, 0.147]
	TimeVAE	0.192	0.003	[0.190, 0.194]	[0.184, 0.196]
	VRNNGAN	0.123	0.003	[0.121, 0.124]	[0.119, 0.129]

4.1.3 Weather Dataset Bootstrap Summary

Weather data presents the highest bootstrap variability, particularly for TimeVAE in forecasting metrics (RMSE std: 0.030), reflecting the dataset's inherent complexity.

Table 5: Bootstrap Summary Statistics for Weather Dataset (15 runs per model)

Metric	Model	Mean	Std Dev	95% CI	Range
KL Divergence	TimeGAN	8.908	0.061	[8.874, 8.941]	[8.773, 9.009]
	TimeVAE	9.848	0.047	[9.821, 9.874]	[9.752, 9.942]
	VRNNGAN	3.072	0.016	[3.063, 3.081]	[3.051, 3.105]
Wasserstein Distance	TimeGAN	0.169	0.002	[0.167, 0.170]	[0.165, 0.174]
	TimeVAE	0.162	0.001	[0.162, 0.163]	[0.159, 0.164]
	VRNNGAN	0.129	0.003	[0.128, 0.131]	[0.126, 0.135]
RMSE	TimeGAN	0.345	0.004	[0.343, 0.347]	[0.339, 0.355]
	TimeVAE	1.440	0.030	[1.423, 1.457]	[1.386, 1.493]
	VRNNGAN	0.304	0.004	[0.301, 0.306]	[0.296, 0.311]
MAE	TimeGAN	0.211	0.003	[0.210, 0.213]	[0.208, 0.218]
	TimeVAE	0.875	0.026	[0.860, 0.889]	[0.831, 0.920]
	VRNNGAN	0.200	0.003	[0.198, 0.201]	[0.195, 0.204]

4.2 Statistical Analysis (ANOVA)

4.2.1 ANOVA Summary

The two-way ANOVA analysis reveals significant performance differences between generative models across all evaluation metrics. Table 6 demonstrates that model type, dataset characteristics, and their interactions all produce statistically significant effects ($p < 0.001$) on synthetic data quality.

The significant main effects confirm that TimeGAN, TimeVAE, and VRNNGAN generate distinctly different synthetic data quality outcomes, with performance variations that are consistent across the four metrics examined. Similarly, the three datasets—electricity, exchange, and weather—present varying challenges for synthetic data generation, indicating that certain time series characteristics inherently affect model performance.

Most importantly, the significant Model $\times$ Dataset interactions across all metrics indicate that model performance is context-dependent rather than universal. This finding

demonstrates that the relative effectiveness of each generative approach varies substantially depending on the specific characteristics of the time series data being modeled. The interaction effects provide statistical justification for conducting dataset-specific analyses rather than relying on aggregated performance comparisons, which could obscure important context-dependent patterns in model effectiveness.

Table 6: Two-way ANOVA summary for synthetic data quality metrics

Metric	**Model F-statistic**	**Dataset F-statistic**	**Model × Dataset F-statistic**	**Significance ($p < 0.001$)**
KL Divergence	300,802	31,336	18,227	***
Wasserstein Distance	115	2,098	3,805	***
RMSE	20,828	32,897	18,110	***
MAE	12,066	14,375	8,758	***

4.2.2 Model × Dataset Interaction

The significant Model × Dataset interactions identified in the ANOVA analysis are visualized through interaction plots that demonstrate how model performance varies across different dataset contexts. These plots reveal non-parallel lines, confirming that the relative effectiveness of each generative model changes substantially depending on the specific characteristics of the time series data being modeled.

Figure 1 illustrates the interaction effects for KL divergence, showing the most pronounced performance variations across datasets. The dramatic convergence of all three dataset lines at VRNNGAN demonstrates the hybrid model's universal effectiveness in preserving statistical fidelity, while the divergent patterns for TimeGAN and TimeVAE across datasets highlight their context-dependent performance.

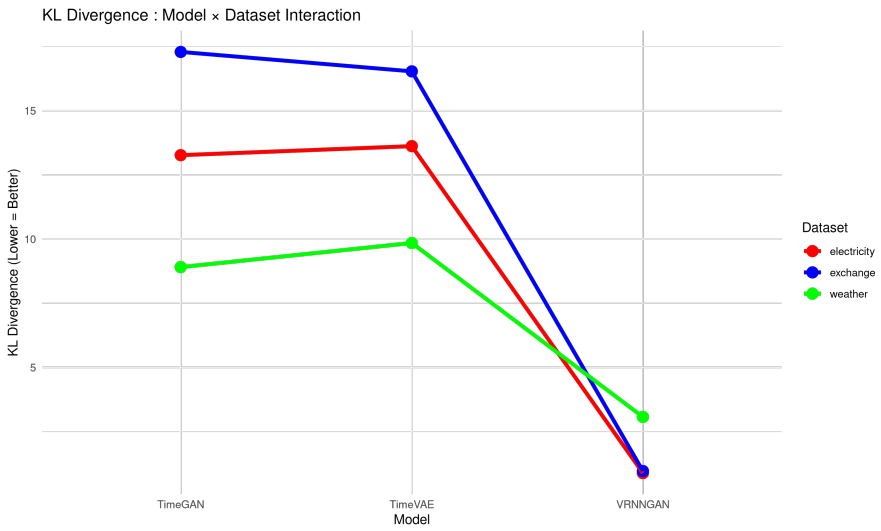

Figure 1: Model × Dataset Interaction Effects for KL Divergence

The Wasserstein distance interactions, shown in Figure 2, reveal different patterns where TimeVAE demonstrates superior performance specifically for electricity data, while TimeGAN shows advantages for exchange data. These crossing lines exemplify how geometric similarity preservation varies significantly by dataset characteristics.

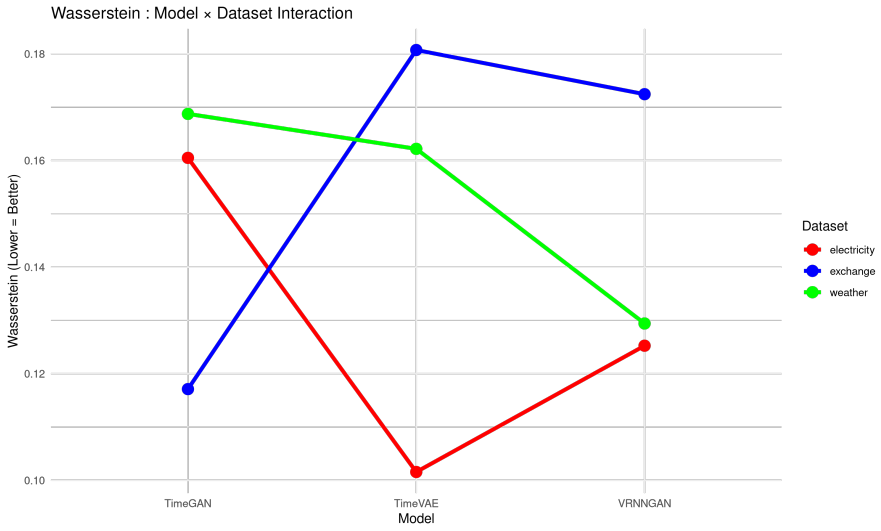

Figure 2: Model × Dataset Interaction Effects for Wasserstein Distance

Figures 3 and 4 display the forecasting performance interactions, where weather data presents unique challenges for all models, particularly affecting TimeVAE performance. The near-parallel lines for electricity and exchange data in both RMSE and MAE suggest more consistent relative model performance for these forecasting metrics across these datasets.

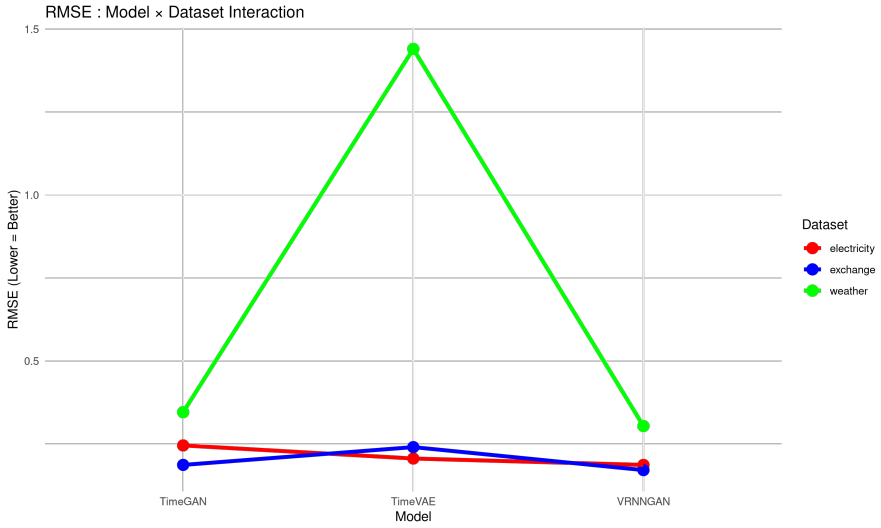

Figure 3: Model × Dataset Interaction Effects for RMSE

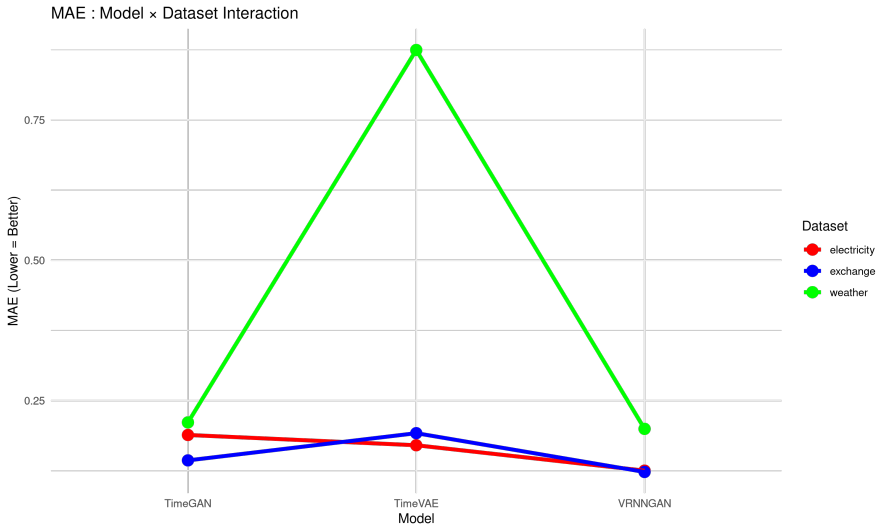

Figure 4: Model × Dataset Interaction Rffects for MAE

These interaction patterns reinforce the ANOVA findings, highlighting that model selection depends on the specific characteristics of the target dataset. The varying interaction strengths across metrics further indicate that different aspects of synthetic data quality, statistical fidelity versus forecasting utility, are affected differently by dataset characteristics. To identify the specific sources of these differences, Tukey HSD post-hoc analysis revealed distinct performance groupings across all metrics, with VRNNGAN consistently forming its own superior performance group while TimeGAN and TimeVAE showed varying relative positions depending on the evaluation context. These statistical foundations necessitate the dataset-specific Shapley analysis that follows to provide

interpretable attribution of each model's contribution to synthetic data quality.

4.2.3 Tukey-HSD Post-Hoc Results

A closer examination of the post-hoc results illustrates these performance groupings in more detail. For KL divergence results in Table 7, VRNNGAN demonstrates substantially superior performance with a mean value of 1.64, forming its own distinct group (c). TimeGAN and TimeVAE show relatively similar but significantly different performance levels, with TimeGAN (13.16, group a) slightly outperforming TimeVAE (13.33, group b). This pattern indicates that while both individual models struggle with statistical fidelity compared to the hybrid approach, TimeGAN maintains marginally better distributional similarity to the real reference data.

Table 7: Tukey HSD Post-Hoc analysis for KL Divergence

Model	**Mean Value**	**Tukey Group**
VRNNGAN	1.6399	c
TimeGAN	13.1554	a
TimeVAE	13.3337	b

The Wasserstein distance results in Table 8 reveal a different performance ranking, where all three models form distinct statistical groups despite relatively small absolute differences. VRNNGAN again achieves the best performance (0.142, group c), followed by TimeVAE (0.148, group b) and TimeGAN (0.149, group a). The closer clustering of values suggests that geometric similarity preservation shows less dramatic differences between models compared to statistical divergence measures.

Table 8: Tukey HSD Post-Hoc Analysis for Wasserstein Distance

Model	**Mean Value**	**Tukey Group**
VRNNGAN	0.1424	c
TimeVAE	0.1481	b
TimeGAN	0.1488	a

For forecasting performance metrics, the post-hoc analysis demonstrates more pronounced differences between models. In RMSE evaluation (Table 9), VRNNGAN main-

tains its superior performance (0.220, group c), while TimeGAN (0.259, group a) significantly outperforms TimeVAE (0.629, group b). This substantial gap indicates that TimeVAE's approach to latent space modeling may not effectively capture the temporal dependencies necessary for accurate forecasting. Similarly, MAE results (Table 10) confirm this pattern, with VRNNGAN (0.149, group c) leading, followed by TimeGAN (0.181, group a), and TimeVAE showing notably higher error rates (0.412, group b).

Table 9: Tukey HSD Post-Hoc Analysis for RMSE

Model	**Mean Value**	**Tukey Group**
VRNNGAN	0.2203	c
TimeGAN	0.2591	a
TimeVAE	0.6287	b

Table 10: Tukey HSD Post-Hoc Analysis for MAE

Model	**Mean Value**	**Tukey Group**
VRNNGAN	0.1492	c
TimeGAN	0.1811	a
TimeVAE	0.4123	b

The consistent emergence of three distinct statistical groups across all metrics confirms that each model contributes unique characteristics to synthetic data generation. VRNNGAN's consistent placement in the superior performance group (c) across all metrics validates the effectiveness of hybrid architectures in balancing multiple aspects of data quality. The varying relative positions of TimeGAN and TimeVAE across different metrics highlight the trade-offs inherent in different generative approaches, with TimeGAN showing particular strength in forecasting tasks and TimeVAE demonstrating competitive performance in certain geometric similarity measures.

4.3 Shapley Value Analysis

The Shapley analysis reveals the individual contributions of each model component to the overall performance across different metrics and datasets. For each metric evaluated, the analysis examines the marginal contributions and interactions between model

components.

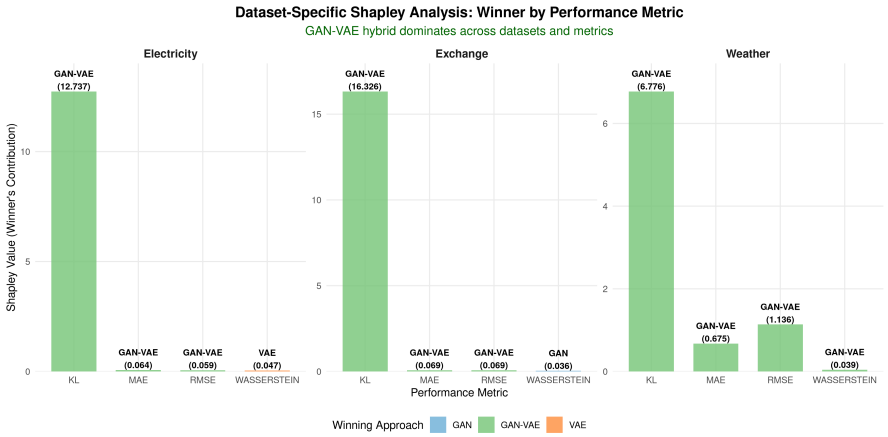

Figure 5: Shapley Analysis Summary Plot

The comprehensive analysis across datasets reveals distinct patterns in model effectiveness. VRNNGAN demonstrates consistent superiority across most metrics and datasets, with particularly strong performance in statistical fidelity measures (KL divergence) and forecasting accuracy (RMSE, MAE). However, individual models show dataset-specific advantages: TimeVAE excels in geometric similarity preservation for electricity data, while TimeGAN demonstrates superior Wasserstein distance performance for exchange data.

4.3.1 Electricity Dataset Results

The electricity dataset demonstrates VRNNGAN's strong performance across most evaluation metrics, with the notable exception of Wasserstein distance where TimeVAE shows superior geometric similarity preservation. Table 11 shows that VRNNGAN achieves the highest Shapley values for KL divergence (12.737), RMSE (0.059), and MAE (0.064), indicating consistent effectiveness in both statistical fidelity and forecasting utility.

TimeGAN's negative Shapley value (-0.012) for Wasserstein distance indicates that TimeGAN reduces geometric similarity in electricity data generation. This negative contribution suggests performance degradation when TimeGAN's adversarial component is included for this specific metric-dataset combination.

Table 11: Shapley value contributions for the electricity dataset

Approach	KL Divergence	Wasserstein	RMSE	MAE
TimeGAN	6.544	-0.012	0.010	0.023
TimeVAE	6.193	**0.047**	0.050	0.041
VRNNGAN	**12.737**	0.035	**0.059**	**0.064**

Note: Bold indicates highest Shapley value (best contributor) per metric.

4.3.2 Exchange Dataset Results

The exchange dataset analysis reveals VRNNGAN's dominance across most metrics, with notable exceptions in geometric similarity preservation. Table 12 demonstrates that for KL divergence, VRNNGAN achieves the highest Shapley value (16.326), indicating exceptional statistical fidelity preservation. However, TimeGAN demonstrates superior performance for Wasserstein distance (0.036), suggesting better geometric similarity modeling for this specific financial dataset context.

The emergence of TimeVAE's negative Shapley value (-0.028) for Wasserstein distance indicates that TimeVAE reduces geometric similarity in exchange data generation. This negative contribution suggests performance degradation when TimeVAE's variational component is included for this specific metric-dataset combination.

Table 12: Shapley value contributions for the exchange dataset

Approach	KL Divergence	Wasserstein	RMSE	MAE
TimeGAN	7.785	**0.036**	0.061	0.059
TimeVAE	8.542	-0.028	0.008	0.010
VRNNGAN	**16.326**	0.008	**0.069**	**0.069**

Note: Bold indicates highest Shapley value (best contributor) per metric.

4.3.3 Weather Dataset Results

The weather dataset presents the most challenging context for synthetic data generation, evidenced by the emergence of negative Shapley contributions. Table 13 reveals that VRNNGAN maintains superior performance across all metrics, with particularly strong contributions for KL divergence (6.776).

Table 13: Shapley value contributions for the weather dataset

Approach	KL Divergence	Wasserstein	RMSE	MAE
TimeGAN	3.858	0.016	1.116	0.669
TimeVAE	2.918	0.023	0.21	0.006
VRNNGAN	**6.776**	**0.039**	**1.137**	**0.675**

Note: Bold indicates highest Shapley value (best contributor) per metric.

4.4 Discussion

The statistical analysis reveals a coherent pattern of findings that illuminate both the individual strengths of generative models and their complex interactions with dataset characteristics. The significant main effects identified through ANOVA (all F-statistics > 100, $p < 0.001$) combined with the Tukey HSD post-hoc results establish VRNNGAN's consistent superiority across evaluation metrics, while the substantial Model $\times$ Dataset interactions demonstrate that synthetic data generation effectiveness is fundamentally context-dependent rather than universal.

Hybrid Architecture Synergy and Performance Patterns

This statistical foundation provides empirical support for the Shapley value analysis, which reveals that VRNNGAN's advantages stem from genuine architectural synergy rather than merely additive benefits of combining VAE and GAN approaches. The hybrid model's KL divergence Shapley values ranging from 6.776 to 16.326 compared to individual models' contributions of 2.918 to 8.542 demonstrate substantial improvements in statistical fidelity preservation. Similarly, VRNNGAN consistently achieves the highest Shapley values for forecasting utility metrics, with RMSE contributions ranging from 0.059 to 1.137 and MAE contributions from 0.064 to 0.675, confirming enhanced practical utility across all datasets.

The consistent superiority of the hybrid architecture extends beyond the comparative advantages identified in previous literature. While Yoon et al. (2019) focused on demonstrating TimeGAN's effectiveness against other GAN variants, and Desai et al. (2021) showed TimeVAE's competitive performance against TimeGAN in specific contexts, this study provides the first systematic evidence that combining the GAN and

VAE approaches yields synergistic benefits. The Shapley analysis framework enables quantification of how GAN and VAE components contribute to overall performance, revealing that VRNNGAN's success stems from architectural complementarity rather than simple performance averaging.

Context-Dependent Model Effectiveness

Notably, while VRNNGAN demonstrates consistent superiority across most evaluation contexts, winning 10 out of 12 metric-dataset combinations in the Shapley analysis, the relative rankings between TimeGAN and TimeVAE are not fixed and vary considerably depending on the evaluation context. This finding reinforces the context-dependent nature of generative model effectiveness and underscores the importance of comprehensive evaluation frameworks rather than relying on single-context comparisons.

Most notably, the Wasserstein distance metric reveals the clearest evidence of context-dependent model effectiveness, with each dataset favoring a different generative approach for geometric similarity preservation. For electricity data, TimeVAE achieves the highest Shapley value (0.047), demonstrating the effectiveness of variational approaches for structured temporal patterns with regular load distributions. In contrast, exchange data favors TimeGAN with the highest Wasserstein Shapley value (0.036), while TimeVAE shows negative contribution (-0.028), indicating that adversarial training better captures the stochastic volatility and irregular patterns characteristic of financial time series. Weather data demonstrates VRNNGAN's superiority (0.039), completing the pattern where each dataset-metric combination reveals distinct architectural advantages.

Model Limitations and Negative Contributions

The emergence of negative Shapley values provides novel insights into model limitations under specific contexts. TimeGAN's negative Wasserstein contribution (-0.012) for electricity data and TimeVAE's negative contribution (-0.028) for exchange data reveal context-specific performance limitations that extend beyond simple underperformance. These negative contributions indicate that certain architectural approaches can actually detract from overall synthetic data quality when applied to incompatible temporal patterns. This represents a fundamental shift from viewing model selection as choosing the best performer to understanding when certain approaches should be explicitly avoided

for specific data contexts.

Methodological Contributions and Framework Advancement

The application of Shapley value analysis to generative model evaluation represents a significant methodological advancement that addresses evaluation challenges identified in recent literature reviews by Goyal and Mahmoud (2024). The framework successfully provides interpretable attribution of model contributions while maintaining mathematical rigor through fair allocation principles. The bootstrap methodology with 15 iterations per model-dataset combination provided sufficient statistical power for robust evaluation while revealing consistent performance patterns that inform practical model selection decisions.

Practical Implications and Implementation Guidance

The findings provide actionable guidance for practitioners implementing synthetic time series data solutions. VRNNGAN demonstrates consistent superiority across statistical fidelity and forecasting metrics, winning 10 out of 12 total metric-dataset combinations, suggesting that hybrid architectures should be prioritized for applications requiring balanced performance across multiple quality dimensions. However, the Wasserstein distance results reveal important exceptions: TimeVAE achieves optimal geometric similarity preservation for electricity data (Shapley value: 0.047), TimeGAN excels for exchange data (0.036), while VRNNGAN dominates weather data (0.039).

CHAPTER 5

CONCLUSION AND FUTURE WORKS

This study explores the application of Shapley value analysis to evaluate generative models for synthetic time series data generation, focusing on the balance between fidelity and usability metrics. The research systematically compared TimeGAN, TimeVAE, and VRNNGAN across three diverse datasets to determine optimal model selection strategies for synthetic data applications. The methodology employed a comprehensive evaluation framework that integrated statistical fidelity measures and forecasting utility metrics within a Shapley-based analysis framework. Bootstrap validation provided statistical robustness, while two-way ANOVA analysis confirmed significant model-dataset interactions across all evaluation metrics. Tukey HSD post-hoc testing identified specific pairwise differences between models, revealing distinct performance groupings that informed the subsequent Shapley value attribution analysis.

The results demonstrate VRNNGAN's consistent superiority across all evaluation scenarios. The hybrid model achieved the highest performance in both distributional similarity and forecasting accuracy, with substantial improvements over individual model approaches. Model effectiveness varied considerably across different time series characteristics, with hybrid advantages ranging significantly depending on context.

The emergence of negative Shapley values provided novel insights into architectural limitations, demonstrating that certain model-dataset conbinations can actively degrade synthetic data quality rather than simply underperforming. This finding fundamentally reframes model selection from identifying the best performed to understanding when specific approaches should be avoided for particular temporal contexts.

The Shapley value framework represents a significant methodological advancement in generative model evaluation, addressing longstanding challenges in interpretable performance attribution. Unlike traditional comparative approaches that provide aggregate performance rankings, this framework enables practitioners to understand why certain models excel in specific contexts and how architectural components contribute to overall synthetic data quality. The mathematical rigor of fair allocation principles ensures that performance attribution remains consistent and interpretable across diverse evaluation scenarios.

The study's limitations include its focus on forecasting tasks using LSTM networks, which may not generalize to other machine learning applications such as classification or anomaly detection. The deliberate exclusion of privacy metrics represents a significant limitation given the critical importance of privacy preservation in synthetic data applications. Additionally, the evaluation scope was constrained to time series data and three specific datasets, which may not capture the full diversity of real-world data characteristics.

Due to the lack of comprehensive privacy evaluation in this framework, future scholars can extend the Shapley-based analysis to incorporate differential privacy metrics, creating a three-dimensional evaluation ecosystem that addresses fidelity, usability, and privacy simultaneously. Future research should explore the framework's applicability to diverse machine learning tasks beyond forecasting and investigate the specific data characteristics that drive the observed model-dataset interactions. Additionally, scaling studies with larger datasets and emerging generative architectures would further validate the framework's broader applicability.

DECLARATION OF CONFLICTS OF INTEREST

The author declares that there is no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY

The datasets used in this study were sourced from PapersWithCode, a publicly accessible repository of machine learning datasets and benchmarks. The three time series datasets employed in this research are available at the following locations:

- **Electricity Transformer Temperature (ETTh1):** Available at https://paperswithcode.com/dataset/etth1-192
- **Exchange Rate:** Available at https://paperswithcode.com/dataset/echange
- **Weather:** Available at https://paperswithcode.com/dataset/weather-ltsf

Access to these datasets is subject to the terms and conditions of PapersWithCode and the original data providers. All datasets are publicly available for research purposes. The preprocessing scripts and experimental code used in this study can be made available upon reasonable request to support reproducibility of the findings.

REFERENCES

Bauer, A., Trapp, S., Stenger, M., Leppich, R., Kounev, S., Leznik, M., Chard, K., & Foster, I. (2024, February). Comprehensive Exploration of Synthetic Data Generation: A Survey [arXiv:2401.02524]. https://doi.org/10.48550/arXiv.2401.02524

Caliskan, H., Yayla, O. F., & Genc, Y. (2023). A Comparative Analysis of Synthetic Data Generation with VAE and CTGAN Models on Financial Credit Loan Offer Data [ISSN: 2521-1641]. *2023 8th International Conference on Computer Science and Engineering (UBMK)*, 212–217. https://doi.org/10.1109/UBMK59864.2023.10286762

Carvajal-Patiño, D., & Ramos-Pollán, R. (2022). Synthetic data generation with deep generative models to enhance predictive tasks in trading strategies. *Research in International Business and Finance*, *62*, 101747. https://doi.org/10.1016/j.ribaf.2022.101747

Chundawat, V. S., Tarun, A. K., Mandal, M., Lahoti, M., & Narang, P. (2024, June). TabSynDex: A Universal Metric for Robust Evaluation of Synthetic Tabular Data [arXiv:2207.05295]. https://doi.org/10.48550/arXiv.2207.05295

Cullen, D., Halladay, J., Briner, N., Basnet, R., Bergen, J., & Doleck, T. (2022). Evaluation of Synthetic Data Generation Techniques in the Domain of Anonymous Traffic Classification. *IEEE Access*, *10*, 129612–129625. https://doi.org/10.1109/ACCESS.2022.3228507

Desai, A., Freeman, C., Wang, Z., & Beaver, I. (2021, December). TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation [arXiv:2111.08095]. https://doi.org/10.48550/arXiv.2111.08095

Duddu, V., Szyller, S., & Asokan, N. (2022, September). SHAPr: An Efficient and Versatile Membership Privacy Risk Metric for Machine Learning [arXiv:2112.02230]. https://doi.org/10.48550/arXiv.2112.02230

El Emam, K. (2020). Seven Ways to Evaluate the Utility of Synthetic Data. *IEEE Security & Privacy*, *18*(4), 56–59. https://doi.org/10.1109/MSEC.2020.2992821

Emam, K. E., Mosquera, L., & Hoptroff, R. (2020, May). *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data* [Google-Books-ID: XWnnDwAAQBAJ]. "O'Reilly Media, Inc.".

Figueira, A., & Vaz, B. (2022). Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics*, *10*(15), 2733. https://doi.org/10.3390/math10152733

Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: A literature review. *Journal of Big Data*, *10*(1), 115. https://doi.org/10.1186/s40537-023-00792-7

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, *20*(1), 108. https://doi.org/10.1186/s12874-020-00977-1

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014, June). Generative Adversarial Networks [arXiv:1406.2661]. https://doi.org/10.48550/arXiv.1406.2661

Goyal, M., & Mahmoud, Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics*, *13*(17), 3509. https://doi.org/10.3390/electronics13173509

Habiba, M., Borphy, E., Pearlmutter, B. A., & Ward, T. (2021). ECG Synthesis with Neural ODE and GAN Models. *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, 1–6. https://doi.org/10.1109/ICECET52533.2021.9698702

Hazra, D., Shafqat, W., & Byun, Y.-C. (2022). Generating synthetic data to reduce prediction error of energy consumption. *Comput. Mater. Contin.*, *70*(2), 3151–3167.

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, *493*, 28–45. https://doi.org/https://doi.org/10.1016/j.neucom.2022.04.053

Hudovernik, V., Jurkovič, M., & Štrumbelj, E. (2024, October). Benchmarking the Fidelity and Utility of Synthetic Relational Data [arXiv:2410.03411 [cs]]. https://doi.org/10.48550/arXiv.2410.03411

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022, May). Synthetic Data – what, why and how? [arXiv:2205.03257]. https://doi.org/10.48550/arXiv.2205.03257

Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Kulynych, B., Prasser, F., & Raisaro, J. L. (2024, August). A Scoping Review of Privacy and Utility Metrics in Medical Synthetic Data. https://doi.org/10.1101/2023.11.28.23299124

Kingma, D. P., & Welling, M. (2022, December). Auto-Encoding Variational Bayes [arXiv:1312.6114]. https://doi.org/10.48550/arXiv.1312.6114

Kondratyev, O., & Schwarz, C. (2019, May). The Market Generator. https://doi.org/10.2139/ssrn.3384948

Lee, J. (2022). *VRNNGAN: A Recurrent VAE-GAN Framework for Synthetic Time-Series Generation* [Master's thesis, University of Toronto]. https://utoronto.scholaris.ca/server/api/core/bitstreams/be516135-4c85-4d0b-9970-be281f59388b/content

Li, W., Ding, W., Sadasivam, R., Cui, X., & Chen, P. (2019). His-GAN: A histogram-based GAN model to improve data generation quality. *Neural Networks*, *119*, 31–45. https://doi.org/10.1016/j.neunet.2019.07.001

Lin, L., Li, Z., Li, R., Li, X., & Gao, J. (2023, May). Diffusion Models for Time Series Applications: A Survey [arXiv:2305.00624]. https://doi.org/10.48550/arXiv.2305.00624

Loni, M., Poursalim, F., Asadi, M., & Gharehbaghi, A. (2025). A review on generative AI models for synthetic medical text, time series, and longitudinal data. *npj Digital Medicine*, *8*(1), 281. https://doi.org/10.1038/s41746-024-01409-w

Lu, Y., Shen, M., Wang, H., Wang, X., Rechem, C. v., Fu, T., & Wei, W. (2024, June). Machine Learning for Synthetic Data Generation: A Review [arXiv:2302.04062]. https://doi.org/10.48550/arXiv.2302.04062

Mendelevitch, O., & Lesh, M. D. (2021, June). Fidelity and Privacy of Synthetic Medical Data [arXiv:2101.08658 [cs]]. https://doi.org/10.48550/arXiv.2101.08658

Nematholahy, A. (2020). Recommender system with RBMs. Retrieved December 13, 2024, from https://kaggle.com/code/alirezanematolahy/recommender-system-with-rbms

Nikolenko, S. I. (2019, September). Synthetic Data for Deep Learning [arXiv:1909.11512]. https://doi.org/10.48550/arXiv.1909.11512

Niu, Z., Yu, K., & Wu, X. (2020). LSTM-Based VAE-GAN for Time-Series Anomaly Detection. *Sensors*, *20*(13), 3738. https://doi.org/10.3390/s20133738

Osorio-Marulanda, P. A., Epelde, G., Hernandez, M., Isasa, I., Reyes, N. M., & Iraola, A. B. (2024). Privacy Mechanisms and Evaluation Metrics for Synthetic Data Generation: A Systematic Review. *IEEE Access*, *12*, 88048–88074. https://doi.org/10.1109/ACCESS.2024.3417608

Prihatno, A. T., Nurcahyanto, H., Ahmed, M. F., Rahman, M. H., Alam, M. M., & Jang, Y. M. (2021). Forecasting PM2.5 concentration using a Single-Dense layer BiLSTM method. *Electronics (Basel)*, *10*(15), 1808.

Qian, Z., Cebere, B.-C., & van der Schaar, M. (2023). Synthcity: Facilitating innovative use cases of synthetic data in different data modalities.

Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted Boltzmann machines for collaborative filtering. *Proceedings of the 24th international conference on Machine learning*, 791–798. https://doi.org/10.1145/1273496.1273596

Shahul Hameed, M. A., Qureshi, A. M., & Kaushik, A. (2024). Bias Mitigation via Synthetic Data Generation: A Review. *Electronics*, *13*(19), 3909. https://doi.org/10.3390/electronics13193909

Shapley, L. S. (1952). *A value for n-person games*. RAND Corporation. https://doi.org/10.7249/P0295

Skandarani, Y., Jodoin, P.-M., & Lalande, A. (2021, July). GANs for Medical Image Synthesis: An Empirical Study [arXiv:2105.05318]. https://doi.org/10.48550/arXiv.2105.05318

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015, November). Deep Unsupervised Learning using Nonequilibrium Thermodynamics [arXiv:1503.03585]. https://doi.org/10.48550/arXiv.1503.03585

Yoon, J., Jarrett, D., & van der Schaar, M. (2019, December). Time-series generative adversarial networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (pp. 5508–5518). Curran Associates Inc.

You, Z., Zhong, Y., Bao, F., Sun, J., Li, C., & Zhu, J. (2023, October). Diffusion Models and Semi-Supervised Learners Benefit Mutually with Few Labels [arXiv:2302.10586]. https://doi.org/10.48550/arXiv.2302.10586

Zhu, J. (2024). Synthetic data generation by diffusion models. *National Science Review*, *11*(8), nwae276. https://doi.org/10.1093/nsr/nwae276

Zia, M., Frazier, S., & Nazaripouya, H. (2023). Synthetic Agricultural Load Data Generation Using TimeGANs [ISSN: 2833-003X]. *2023 North American Power Symposium (NAPS)*, 1–6. https://doi.org/10.1109/NAPS58826.2023.10318596