Austin Journal of Microbiology

Open Access

Austin Publishing Group

Review Article

Evaluating the Efficacy of T-GAN and W-GAN Augmented Data in Machine Learning Models

Murad Ali Khan*

Department of Computer Engineering, Jeju National University, Jeju 63243, Republic of Korea

***Corresponding author: Murad Ali Khan**

Department of Computer Engineering, Jeju National University, Jeju 63243, Republic of Korea.

Email: muradali@stu.jejunu.ac.kr

Received: May 30, 2024
Accepted: June 13, 2024
Published: June 20, 2024

Abstract

This paper presents a comparative analysis of the performance of Time-GAN (T-GAN) and Wasserstein-GAN (W-GAN) augmented data using various machine learning models, including Extra Trees, XGBoost, CatBoost, and Light GBM. Utilizing multiple metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2, Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE), the study aims to determine which GAN technique produces the most effective synthetic data for enhancing model performance. The results indicate that T-GAN augmented data generally achieves better performance metrics, particularly when used with the Extra Trees model.

Keywords: Augmentation; Synthetic Data; GAN; Machine Learning; Artificial Intelligence; Data extension; Clinical data; Data Privacy; Regulatory Compliance

Introduction

The advent of machine learning and its application across various domains has necessitated the development of robust methods to generate and utilize synthetic data effectively. GANs have emerged as a prominent solution for data augmentation, particularly in areas plagued by data scarcity or privacy concerns. This paper focuses on two sophisticated GAN variants, T-GAN) and W-GAN, which have been tailored to enhance the realism and utility of synthetic data in predictive modeling. By leveraging these technologies, our study aims to evaluate and compare their efficacy in generating data that not only mirrors real-world distributions but also effectively enhances the performance of machine learning models. Recent studies like [1,2] provide foundational support, showcasing the advancements in GAN architectures and their application in complex datasets.

The use of T-GAN and W-GAN in this context is particularly relevant due to their distinct approaches to handling the inherent challenges of data generation, such as maintaining temporal coherence in time-series data and addressing the mode collapse problem in training GANs. Through a detailed analysis using various metrics and machine learning models, this study seeks to identify which GAN methodology better supports data-driven decision-making in predictive analytics. Insights from recent articles [3-5] contribute to the understanding of how synthetic data influences model accuracy and training efficiency, guiding this paper's exploration of GAN utility in sports analytics.

Related Work

The application of GANs for synthetic data generation has been extensively documented across multiple fields, including healthcare, finance, and sports analytics. Researchers have increasingly turned to these networks to address issues of data limitation and improve model training under constrained conditions. Pioneering works by Goodfellow et al. introduced the foundational GAN framework, which has since been adapted and refined through numerous studies. The integration of GANs into complex applications, such as those discussed by [6,7], has further established their critical role in data augmentation practices across industries.

Further advancing the discussion, [8] explored the practical applications of GANs in engineering, providing critical insights into their potential to replicate and extend real-world data scenarios accurately. These studies collectively underscore the versatility and adaptability of GANs, setting a precedent for their use in enhancing datasets for predictive modeling. Additionally, recent publications by [9,10] highlight innovative uses of GANs in creating realistic synthetic datasets for training algorithms under resource constraints, emphasizing their importance in contemporary data science. This paper builds upon these insights by focusing specifically on T-GAN and W-GAN, analyzing their unique contributions and effectiveness in generating high-quality synthetic data. Through a comparative analysis, this work contributes to the ongoing dialogue about the best

Citation: Khan MA. Evaluating the Efficacy of T-GAN and W-GAN Augmented Data in Machine Learning Models. Austin J Microbiol. 2024; 9(2): 1052.

practices for employing GANs in complex, multi-indexed data environments such as athlete performance metrics.

Proposed Framework

This section, outlines the systematic approach taken to enhance the quality and quantity of the dataset used for building predictive models. It comprehensively details the methodologies for preprocessing raw, multi-source data into a clean, normalized, and reliable format. Moreover, it introduces the innovative use of Time-GAN and W-GAN for data augmentation, aiming to enrich the dataset with realistic, synthetic samples. These efforts are critical for overcoming limitations associated with small datasets and ensuring robust model training. Finally, the section discusses the evaluation of the augmented data using advanced machine learning algorithms like Extra Trees, XGBoost, CatBoost, and LightGBM to assess the efficacy of the data augmentation techniques employed, ensuring the models are both accurate and scalable.

Data Preprocessing

To address the challenge of multiple-source data collection, which often results in non-uniform datasets, we propose a comprehensive preprocessing model. This model is designed to transform raw, disparate data into a consistent and reliable format suitable for subsequent analysis, as shown in Figure 1. The preprocessing steps include:

- **Data Cleaning:** This involves the removal of duplicate records and irrelevant features, thereby reducing computational and storage demands. Such cleaning not only enhances the efficiency of the model but also improves the accuracy of the predictions by focusing on pertinent data.

- **Handling Missing Values**: To tackle the issue of missing data, our model utilizes zero filling and K-Nearest Neighbors (KNN) imputation methods. These techniques help in maintaining the integrity of the dataset without compromising the quality of the data.

- **Data Aggregation**: Given the temporal nature of our data, specifically concerning athlete performance metrics collected over an 11-day period (from June 10, 2021, to June 20, 2021), aggregation is performed to consolidate daily measures into a single, coherent dataset.

- **Normalization**: To ensure that the attribute values are on a comparable scale, Min-Max normalization is applied, transforming the data into a uniform range of [1]. This step is crucial for models that are sensitive to the scale of input features.

Data Augmentation

To augment the limited dataset and generate a more robust training set, we employ two Generative Adversarial Networks (GANs):

- **Time-GAN:** This method is particularly suited for extending multi-indexed datasets, such as the time-series data of athletes. Time-GAN respects the temporal correlations within the data, thereby producing synthetic instances that are realistic and time-consistent.

- **Wasserstein GAN (W-GAN):** Known for its stability and effectiveness in generating high-quality samples, W-GAN is used to create additional data points that adhere to the original data distribution. This model addresses the mode collapse issue often encountered in traditional GANs, ensuring diversity in the synthetic data produced.

Figure 2 provides a detailed visual comparison between real and synthetic data across various feature outcomes, reflecting the effectiveness of the data augmentation techniques implemented. Each plot illustrates the performance of specific outcomes, such as controller response metrics and safety indices, highlighting the discrepancies and alignments between the real and synthetic datasets. Notably, synthetic data shows considerable fluctuations across different outcomes, demonstrating variability that is typical of generative models like Time-GAN and W-GAN.

However, in the case of the “strong safety index”, the overlap between real and synthetic values is much closer, suggesting that the augmentation methods are particularly effective in replicating complex, real-world scenarios where safety is a critical measure. This comparative analysis is essential for understanding the strengths and limitations of synthetic data in mimicking real operational conditions.

Model Evaluation

For the evaluation of the performance and reliability of the augmented data produced by CT-GAN and W-GAN, we deploy several advanced machine learning algorithms:

- **Extra Trees:** This ensemble method, known for its high accuracy and ability to run efficiently on large datasets, serves as one of our primary evaluators.

- **XGBoost:** As a highly efficient and scalable end-to-end boosting system, XGBoost is utilized to assess the performance gains from our augmented datasets.

- **CatBoost:** Recognized for its handling of categorical data and robustness against overfitting, CatBoost provides insights into the effectiveness of our data augmentation in varied scenarios.

- **LightGBM:** This gradient boosting framework is particularly advantageous for its speed and efficiency, making it an ideal candidate for evaluating large augmented datasets.

These models help quantify the improvements in predictive accuracy and model robustness afforded by the augmented data, thereby validating the effectiveness of our proposed data augmentation techniques. Table 1 presents the performance metrics of above models when trained using two different augmented data using: T-GAN and W-GAN. The model’s performance is measured using six metrics: MAE, MSE, RMSE, R2, RMSLE, and MAPE.

For the T-GAN augmented data, the Extra Trees model achieves the best performance across almost all metrics, with an MAE of 0.0067, MSE of 0.0008, RMSE of 0.0206, R2 of 0.9913, RMSLE of 0.0078, and MAPE of 0.0044. This indicates that Extra Trees are highly effective when used with T-GAN data, providing the most accurate and reliable predictions. The other models, such as XGBoost, CatBoost, and Light GBM, also perform well but not as consistently across all metrics compared to Extra Trees.

When considering the W-GAN augmented data, CatBoost shows the most favorable results with an MAE of 0.0135, MSE of 0.0009, RMSE of 0.0297, R2 of 0.9759, RMSLE of 0.021, and a higher MAPE of 0.1589. While CatBoost outperforms other models on several metrics, it has a significantly higher MAPE, indicating larger errors in percentage terms compared to other models. Extra Trees and Light GBM also show competitive performance but with slightly higher errors across some metrics.

In conclusion, T-GAN appears to be the better data augmentation technique, particularly when paired with the Extra Trees model, as it consistently achieves lower errors and higher R2 values, indicating better accuracy and fit. W-GAN also shows promise, especially with CatBoost, but the higher MAPE suggests it might not be as reliable for minimizing percentage errors. Therefore, T-GAN with Extra Trees would be the recommended combination for optimal model performance.

In addition, Figure shows a R2 score comparison of the implemented regression models for predicting strong index. The R2 score of Extra Trees Regressor is 0.9913, which indicate that the model significantly performed well compared to the regression models implemented using W-GAN data. In contrast, Light GBM produced a low R2 score of 0.8991.

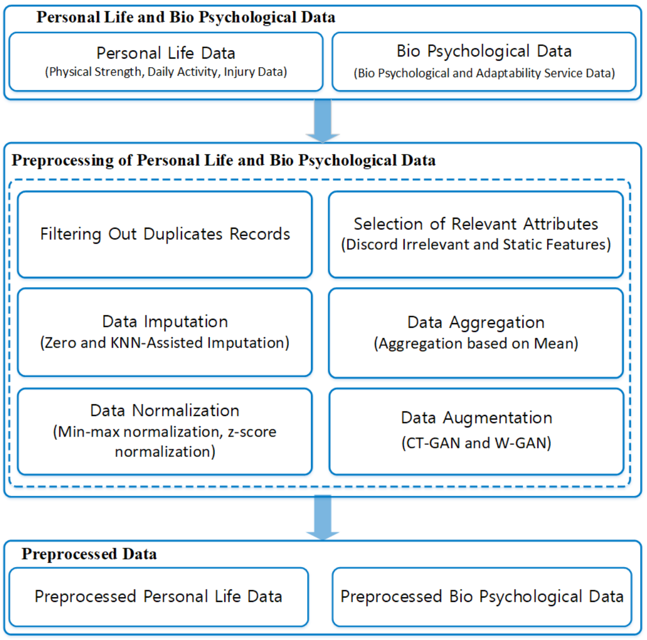
Personal Life and Bio Psychological Data
Personal Life Data
(Physical Strength, Daily Activity, Injury Data)
Bio Psychological Data
(Bio Psychological and Adaptability Service Data)
Preprocessing of Personal Life and Bio Psychological Data
Filtering Out Duplicates Records
Selection of Relevant Attributes
(Discord Irrelevant and Static Features)
Data Imputation
(Zero and KNN-Assisted Imputation)
Data Aggregation
(Aggregation based on Mean)
Data Normalization
(Min-max normalization, z-score normalization)
Data Augmentation
(CT-GAN and W-GAN)
Preprocessed Data
Preprocessed Personal Life Data
Preprocessed Bio Psychological Data

Figure 1: The process of converting initial raw data into machine readable format.

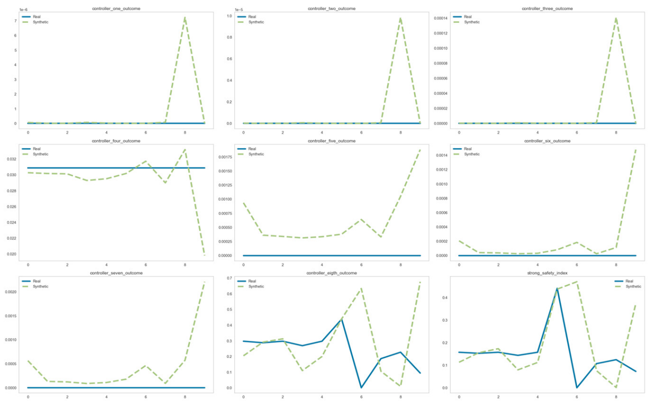

Figure 2: Comparative Analysis of Real and Synthetic Samples.

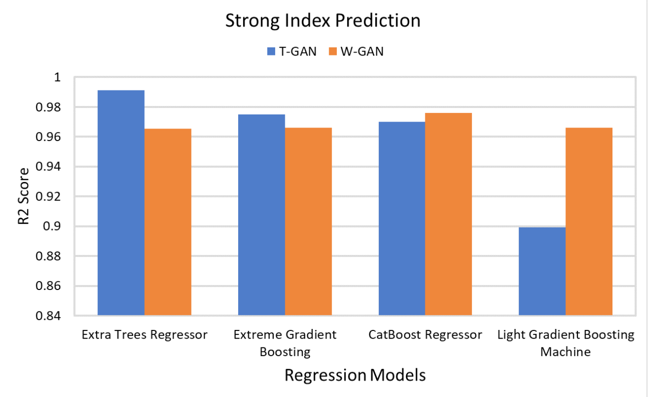

Figure 3: R2 Score Analysis of Regression Models using T-GAN and W-GAN for Predicting Strong Index.

Table 1: Comparative Analysis of Regression Models Performance using Time GAN and W-GAN Data for Strong Index Prediction.

AUG	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
T-GAN	Extra Trees	**0.0067**	**0.0008**	**0.0206**	**0.9913**	**0.0078**	**0.0044**
	XGBoost	0.0104	0.0022	0.0351	0.9749	0.0133	0.0066
	CatBoost	0.0135	0.0027	0.0393	0.9702	0.0154	0.0096
	Light GBM	0.0344	0.0085	0.0796	0.8991	0.0309	0.0235
W-GAN	CatBoost	**0.0135**	**0.0009**	**0.0297**	**0.9759**	**0.021**	0.1589
	Light GBM	0.0152	0.0012	0.0346	0.9675	0.0242	0.1453
	XGBoost	0.0147	0.0013	0.0349	0.966	0.0245	0.1165
	Extra Trees	0.014	0.0013	0.0356	0.9654	0.0248	**0.0737**

Conclusion

The analysis clearly demonstrates that T-GAN augmented data, particularly when analyzed using the Extra Trees model, consistently outperforms W-GAN across most performance metrics. While W-GAN also shows promise, particularly with the CatBoost model, its higher MAPE values suggest it may not always provide the most reliable error minimization. Consequently, T-GAN paired with Extra Trees emerges as the optimal choice for enhancing predictive model performance using synthetic data. This study not only underscores the importance of selecting suitable GAN techniques for data augmentation but also highlights the need for targeted model selection to fully capitalize on the enhanced data quality.

References

1. Peres RS, Azevedo M, Araujo SO, Guedes M, Miranda F, Barata J. Generative adversarial networks for data augmentation in structural adhesive inspection. Applied Sciences. 2021; 11: 3086.

2. Muthukumar P, Zhong J. A stochastic time series model for predicting financial trends using nlp. arXiv preprint. 2021: arXiv:2102.01290.

3. El Emam K, Mosquera L, Fang X, El-Hussuna A. Utility metrics for evaluating synthetic health data generation methods: validation study. JMIR medical informatics. 2022; 10: e35734.

4. Rizvi SKJ, Azad MA, Fraz MM. Spectrum of advancements and developments in multidisciplinary domains for generative adversarial networks (GANs). Archives of Computational Methods in Engineering. 2021; 28: 4503-4521.

5. Jordon, James, Jinsung Yoon, Mihaela Van Der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. International conference on learning representations. 2018.

6. Lamba S, Saini P, Kaur J, Kukreja V. Optimized classification model for plant diseases using generative adversarial networks. Innovations in Systems and Software Engineering. 2023; 19: 103-115.

7. Dogariu Mihai, Stefan LD, Boteanu BA, Lamba C, Kim B, Lonescu B. Generation of realistic synthetic financial time-series. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2022; 18: 1-27.

8. Xu H, Li C, Rahaman MM, Yao Y, Li Z, Zhang J, Kulwa F, et al. An enhanced framework of generative adversarial networks (EF-GANs) for environmental microorganism image augmentation with limited rotation-invariant training data. IEEE Access. 2020; 8: 187455-187469.

9. Al–Qerem A, Ali AM, Attar H, Nashwan S, Qi L, Moghimi MK et al. Synthetic Generation of Multidimensional Data to Improve Classification Model Validity. ACM Journal of Data and Information Quality. 2023; 15: 1-20.

10. Mumuni A, Mumuni F. Data augmentation: A comprehensive survey of modern approaches. Array. 2022; 16: 100258.