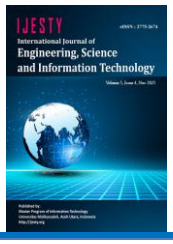
International Journal of Engineering, Science and Information Technology
Volume 5, No. 4 (2025) pp. 268-277
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i4.1442
Research Paper, Short Communication, Review, Technical Paper



Synthetic Data for Business Intelligence: A New Paradigm for Privacy-Preserving Machine Learning in Enterprise Environments

Deep Barot[1]*, Kamal Mohammed Najeeb Shaik[2], Mohammad Mushfiqul Haque Mukit[3], Vinesh Melath[4], Rithesh Nair[5]

[1]Blitz Infocom, United States
[2]Palo Alto Networks Inc, United States
[3]Washington University of Science and Technology, United States
[4]IT Managed Services Genpact, United States
[5]Delivery Management, Cloud Infra and Apps Services (US&C), Unisys Corporation, Albany, NY, Delivery Management, United States

**Corresponding author Email: deeprao0710@hmail.com*

The manuscript was received on 22 February 2025, revised on 15 May 2025, and accepted on 10 August 2025, date of publication 11 November 2025

Abstract

The growing demand for data-driven decision-making in the enterprise context poses a conflict between the utilisation of machine learning (ML) and data privacy. The paper examines the feasibility of using synthetic data to replace actual enterprise data in business intelligence (BI) applications. Synthetic datasets were created using the CTGAN, Variational Autoencoders (VAE), and diffusion models and were successfully assessed in fraud detection and customer segmentation tasks. Empirical findings indicate that XGBoost with synthetic data as training data achieved an accuracy value of 97 percent, with an ROC AUC of 0.94, which is relatively close to the achievable accuracy with real data. CTGAN was found to have high fidelity as the Wasserstein distances were less than 0.15, and the Jensen-Shannon divergence was less than 0.08. The visualisations of dimensionality reductions ensured that the real and synthetic data had a substantial structural similarity. Privacy analyses revealed that the Nearest Neighbour Adversarial Distance (NNAD) scores differed between CTGAN and diffusion models, with values of 0.38 and 0.36, respectively. Corresponding Membership Inference Attack (MIA) success rates were 51-52%, which is significantly lower than the 68% success rate of the anonymised real data. These findings confirm the consideration that synthetic data can maintain analytical value and diminish privacy risks, providing an effective approach to the safe and scalable implementation of ML in businesses.

Keywords: *Synthetic Data, Privacy Risk Mitigation, Business Intelligence, Fidelity Assessment, Generative Models.*

1. Introduction

Enterprise decision-making in the modern data-driven economy relies on the effective strategic application of Business Intelligence (BI) systems, which are based on machine learning (ML). The respective systems feature operational forecasting, customer behaviour modelling, risk management, and market optimisation. With a growing number of applications outperforming classic rule-based systems, business organisations are now widely using machine learning models to automate the extraction of meaning in significant and even challenging data stores [1]. The data that is sensitive, however, and falls under legal, ethical, and regulatory restrictions is the lifeblood of these systems: enterprise data. Whether measuring customer transactions and monetary accounts or individual health communications and behavioural histories, enterprise data sets are full of personally identifiable information (PII) and sensitive attributes.

This increased use of data has drawn increased attention from both regulators and the public [2]. Advances in the realm of global regulating strategies have been exemplified such as the General Data Protection Regulation (GDPR) in Europe, the health-insurance Portability and Accountability Act (HIPAA) in the United States, and the California Consumer Privacy Act (CCPA), which prohibit the collection, storage, and analysis of personal and enterprise data using stringent requirements. Although such frameworks are critical in protecting the privacy of users, they present a significant restriction on the experiments that organisations may undertake, as well as the sharing or even internal use of its data resources. Obedience is mandatory in supervised markets such as finance, medical care, and shipping, and regulations in these areas tend to either lengthen machine learning projects or build suits among lawful threats and accurate forecasting [3].

One of the main difficulties in achieving successful machine learning systems in BI of enterprises is limited access to real-life data due to privacy issues [4]. Such concerns restrict innovation, testing, and the capability to deploy scalable AI solutions. An example would be

that data scientists developing fraud detection models do not have access to the complete transaction logs, or rather, marketing analysts have access only to highly anonymised data that have no real features [5]. Common ways to cover this are either data masking, tokenisation, or aggregation. However, these techniques have the propensity to compromise the quality and design of data, especially multivariate relationships and temporal dependencies, whose intensive use underlies robust machine learning models.

Additionally, the methods of anonymisation are also being criticised by studies that explore the possibility of de-anonymising anonymised data with the help of auxiliary information, as well as re-identification algorithms [6]. Therefore, access to data required for accurate and innovative machine learning is restricted, enabling many organisations to comply yet leaving them in a paradox. This limit insourced group cooperation extends deployment cycles and can result in poorly formed models that are not representative of the real world. It is evident that a more effective and scalable solution is needed, i.e., one that strikes a balance between data-driven innovation and the necessity of privacy protection [7].

This paper has aimed to investigate and confirm the value of synthetic data as a simple and privacy-friendly alternative to real-world data in dealing with machine learning in enterprise BI situations. Synthetic data is specially produced data that is statistically tested to simulate genuine data; however, it lacks any genuine information about users and businesses. The paper analyses the usefulness of synthetic data in various enterprise BI applications, including customer segmentation, fraud detection, and demand forecasting. The central assumption is that synthetic data quality can be used to produce models that perform similarly to those based on real data, while also being regulatory compliant and low-risk in terms of privacy.

Through the latest generative methods, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and tabular diffusion models, this study will develop and test synthetic datasets targeting structured business use cases. Synthetic datasets will be evaluated for their adherence to the original data, their usefulness in training accurate machine learning platforms, and their resistance to the possibilities of reidentification or data leakage. Based on the BankSim synthetic banking transactions dataset available on Kaggle as a foundation, the paper will recreate real-life enterprise conditions to prove the effectiveness of the proposed approach.

The paper presents several new findings related to the topic of privacy-aware machine learning in business environments. First, it offers a comprehensive methodology for generating, evaluating, and deploying synthetic datasets to support business intelligence tasks. This framework achieves a combination of data generation using superior generative models, fidelity testing based on statistical similarity measures, utility testing based on ML performance measurements, and privacy testing based on attack simulation procedures and distance-based risk functions. Second, the paper involves an experimental comparison of models trained on real and synthetic data on the detail level, which is used to prove empirical evidence on the validity of synthetic alternatives. Third, it suggests a list of practical principles for incorporating synthetic data into enterprise ML pipelines. These principles comprise compatibility with MLOps, regulatory alignment, and best practices regarding data governance and auditability. Finally, the study will fill the gap in the analysis between academic progress in generative synthetic data creation and implementation in the real-world BI environment.

The paper is structured as follows. Section 2 presents a literature review on synthetic data generation, its advantages over traditional anonymisation methods, and its applications in business intelligence. Section 3 outlines the methodology employed in this study, including the datasets used, generative models, machine learning models, and evaluation metrics to assess fidelity, utility, and privacy metrics. In Section 4, the results and analysis of the experiments comparing the performance of models trained on real and synthetic data are presented in detail. It also provides visualisations and trade-offs, such as confusion matrices, ROC curves, and distribution similarity plots. Section 5 proposes a scalable architecture for incorporating synthetic data into enterprise ML systems, emphasising its real-world practicality, MLOps friendliness, and compatibility with CI/CD pipelines. Section 6 discusses the ethical, regulatory, and security implications of using synthetic data, explaining that it is at least, and in some instances, more effective than specified in compliance rules while still maintaining analytical capabilities. Last, Section 7 brings the story to its end by summarising findings, the strategic implications of synthetic data on businesses, as well as recommendations for potential future research (such as domain-specific creation of synthetic data and time-series developments).

This organisation's strategy aims to ensure that the paper not only reflects a theoretical contribution but also provides practical suggestions that can be applied by enterprise practitioners seeking to implement privacy-compliant data-driven machine learning models at scale.

2. Literature Review

2.1. Synthetic Data Landscape

Synthetic data development has been closely tied to the complexity of machine learning and regulatory pressure on utilising personal and enterprise data. The initial approaches to synthetic data generation were largely rule-driven, with data generated by domain experts based on their best guesses about the structure of real-world data. In SLA settings of limited scope, the simulators were helpful but failed to accurately represent the relationships within enterprise data that feature complexity and high dimensions [8]. The appearance of probabilistic modelling brought forth additional flexibility; however, these techniques remained limited by poor scalability and accuracy when faced with complex dependence among features.

The modern phase of synthetic data began in 2014 with the introduction of Generative Adversarial Networks (GANs) [9]. GANs introduced a novel model of data generation, relying on a generator-discriminator system to iteratively enhance the generation of synthetic-looking data. Table types of data widely used in enterprise BI have seen considerable success with special-purpose extensions like CTGAN and TableGAN [10]. Simultaneously, Variational Autoencoders (VAEs) presented another algorithm that relies on latent space encoding and probabilistic reconstruction and is especially applicable in controlled data generation. Recently, diffusion models, initially trained on image synthesis, have been adapted for use in structured data synthesis, as they have demonstrated the capacity to handle noisy or unbalanced distributions.

The study has highlighted the fact that synthetic data could be applied as an effective alternative to real-life, highly sensitive data. For example, experimental analyses [11] demonstrated that synthetic data can be made competitive with real data in nearly all classification tasks while providing high privacy assurances. Likewise, the introduction of the GAN-based medGAN model, specifically designed to work with healthcare data, showed the ability of domain-specific generators to balance both effects of fidelity and privacy [12]. All these together form the basis of investigating synthetic data in a high-stakes, privacy-sensitive enterprise context.

2.2. Applications in Enterprise BI

Synthetic data has gained momentum in various impactful business intelligence applications. Customer segmentation is one such field in which organisations aim to segment customers according to purchase patterns, demographic characteristics, or patterns of engagement [13]. Customer-centred traditional clustering and classification models employed in this task are rather data-intensive and are frequently limited by privacy rules. Synthetic data enables businesses to emulate typical customer personas and model-train without using a real identity [14].

Fraud detection is another important use. Banking institutions, online stores, and online services are constantly trying to isolate and avert fraudulent acts in real-time. Nonetheless, fraudulent datasets are sensitive, imbalanced, and have very limited positive examples. The use of GANs or VAE-based synthetic oversampling will not only contribute to solving the issue of class imbalance but also contribute to a safer training process, as the introduction of real financial transactions becomes unnecessary [15]. The BankSim dataset is a typical example of such a synthetic financial transaction dataset that has become a reference for this type of application.

Another area where synthetic data can have a significant impact is demand forecasting. Sales, supply chains, and market trend time-series data enable industries to determine product demand [16] accurately. This information typically contains sensitive business data that is not intended to be shared or analysed within a third-party analytics location. By producing synthetic time-series data that preserves its temporal dependence, organisations can develop forecasting models safely and practically [17]. Similarly, risk modelling in insurance, credit assessment, and operational analysis utilises synthetic data that accurately reflects risk factors, as well as behavioural signals, without compromising the privacy of individuals or institutions involved.

2.3. Privacy Techniques vs Synthetic Generation

Data analysis that preserves the privacy of individuals has long been studied via anonymisation techniques, data masking, and differential privacy. Anonymisation involves stripping out personally identifiable information (PII); however, research has demonstrated that this technique is becoming increasingly susceptible to re-identification, particularly with the involvement of auxiliary datasets [18]. Data masking preserves the field structure while altering relation patterns by replacing sensitive fields with a hash value or another random value. It is more likely to negatively affect model performance in tasks where fine-grained combinations between specific fields are needed.

Differential privacy provides a more formal privacy guarantee by adding noise to data queries or training procedures under control [19]. It is robust in theory but not valuable for practice within a BI environment, as it is complex and reduces the accuracy of the models. By contrast, the structure fidelity can be preserved through synthetic data generation without decoupling from the original records. Synthetic data presents downstream analytics without mentioning or revealing an actual entity when properly generated. In contrast to masking or anonymisation, synthetic data generation is proactive and generative, providing a new data stream in which research and development can occur without regulatory complications [20].

2.4. Research Gap

Along with the increasing body of literature and practical usage examples, a gap remains in the systematic, enterprise-relevant assessment of synthetics. Although several scholarly reports have discussed fidelity or privacy individually, not many measures of all three dimensions, fidelity, utility, and privacy, exist in a single framework that is applicable in enterprise BI [21]. The majority of the available works are either biased to the healthcare industry, too technical, or not aligned with the reality of deployment. There is also a lack of vision on how synthetic data should be integrated into enterprise pipelines, how performance should be monitored, and how regulatory compliance should be ensured over time.

Furthermore, the contextual requirements of applications used in business enterprises, including regulatory auditability, system integration, governance policy compliance, and MLOps compatibility, are rarely discussed in the synthetic data literature [22]. These practical dimensions are essential for application in the real world, which is often a lower priority compared to theoretical novelty. As enterprises view synthetic data as a long-term solution to data accessibility and privacy issues, the lack of a substantial and workable framework poses a significant impediment to its adoption.

This study aims to bridge that gap by providing a comprehensive assessment and implementation plan for synthetic data in enterprise business intelligence. By demonstrating how to generate synthetic data with the real set of generative models and a practical use case analysis of the BankSim dataset, the paper will not only test synthetic data as being statistically similar or equivalent to their real-life counterparts but will also see how it performs when installed in ML models and governed and regulated by the privacy and compliance priorities.

3. Methodology

3.1. Dataset Description

This research utilises the BankSim dataset, a simulated data model based on agent-based modelling that simulates realistic banking payment transaction behaviour. BankSim, a software sourced from Kaggle, simulates the activities of bank customers through the dynamics of agents in a bank's workflow. The data frame contains more than 5,000 records with attributes including transaction type, timestamp, customer ID, merchant ID, amount, fraud flag, and location. The structure resembles real-life financial data substantively and does not contain any personal data; thus, it is the most suitable for conducting privacy-preserving machine learning experiments. Utilising labelled fraudulent transactions enables the implementation of supervised learning models that can be used to detect fraud. Furthermore, the inclusion of a variety of transaction types and customer behaviour in the dataset enables the use of the data to simulate additional business intelligence activities, such as customer segmentation and profiling.

Although the BankSim data itself is synthetic, it can be considered a trustworthy foundation for generating and augmenting synthetic data using superior generative models. The safe use of privacy attacks in this study, which is conducted on controlled synthetic data, is also not compromised by any ethical issues. As part of making data relevant and scalable, redundant fields are dropped, and the data is restructured to fit into the input requirements of both the machine learning and generative modelling frameworks (redundant fields dropped and data rearranged) in future stages.

3.2. Synthetic Data Generation

In this study, the three most advanced generative models (Conditional Tabular GAN (CTGAN), Variational Autoencoder (VAE), and tabular diffusion models) will be adopted to generate synthetic data. These models were identified by their ability to excel at modelling high-dimensional structured data, a significant requirement in enterprise BI environments.

CTGAN is a GAN-specific tabular bundle. It has the capability of dealing with both continuous and categorical variables, as it can model conditional distributions. A mode-specific normalisation method maintains a balance of categories in the generated nominal variables. CTGAN employs a generative discriminator framework, where the generator is trained to produce artificial examples indistinguishable from real data, and the discriminator attempts to distinguish between real and artificial ones.

VAE is another deep probabilistic model that can be used to encode an input into a representation in a latent space and then reconstruct the hidden representation using a decoder. The reconstruction-based approach enables the VAE to generate new data samples that retain the statistical structure of their dataset. In this work, the VAE is parameterised as fully connected networks with Gaussian priors; this configuration makes the transitions between latent spaces smooth, resulting in an improvement in the fidelity of the synthesised samples.

Tabular diffusion models, a recent development in generative modelling, utilise iterative denoising of random noise to synthesise structured datasets in the form of realistic data samples. Perpetuated by their success in image synthesis, these models are extended here to learn tabular data with noise-aware training schedules and feature-wise diffusion transitions. The method is beneficial for producing data with balanced variance and preventing mode collapse, a common problem with GANs.

To train each model, 80 percent of the pre-processed BankSim data is used, and the other 20 percent is reserved as a validation and evaluation dataset. Grid search is used to tune hyperparameters on each model, and early stopping is used to combat overfitting. The output of all models will be a synthetic dataset with the same schema and scale as the original, which will help in downstream machine learning applications and privacy analysis.

3.3. Evaluation Criteria

The synthetically produced data is measured according to three significant aspects- fidelity, utility, and privacy.

Fidelity is the degree to which the synthetic data mimics the distribution of statistical values of the actual set. It is assessed using similarities in distributions, including Wasserstein distance, Jensen-Shannon divergence, and visual inspection, as well as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t-SNE) plots. These steps are taken to preserve the multivariate structure and feature correlation in the synthetic data samples.

The task of utility assessment involves using both the original and synthetic datasets to train and test various machine-learning models. The objective is to understand the extent to which models trained on synthetic data can perform their intended real-world tasks, such as fraud detection. A comparison is made using metrics like accuracy, precision, recall, F1-score, and ROC-AUC. All three classifiers, namely Logistic Regression, Random Forest, and XGBoost, are implemented consistently across the datasets to provide fair analysis. The training of the models is performed using real data, synthetic data, and a hybrid (augmented) dataset to compare performance and generalisation.

Privacy is evaluated using distance-based measures of risk and automated attacks. In particular, the nearest neighbour adversarial distance (NNAD) is calculated to approximate the extent to which synthetic samples resemble any single real record, thereby evaluating the risk of disclosure. Additionally, membership inference attacks (MIA) are simulated to determine whether an adversary can infer whether a given record was included in the training of the generative model. Such privacy tests play a critical role in showing how well the synthetic data is resistant to re-identification and overfitting.

3.4. Machine Learning Models

Utility testing machine learning models, including Logistic Regression, Random Forest, and XGBoost, are characterised by various singular strengths in terms of interpretability, robustness, and performance. Logistic Regression is selected due to its deep interpretability and simplicity as a baseline or benchmark for linear decision boundaries. Random Forest is a tree-based ensemble algorithm that is implemented as such due to its good performance on imbalanced datasets and its ability to capture nonlinear dependencies. XGBoost is a gradient-boosting framework that is applied due to its high performance and negligible scaling costs, especially at the enterprise level.

All models are trained individually with real, synthetic, and hybrid data. Through 5-fold cross-validation, learning rate, maximum depth, the number of estimators, and regularisation terms are optimised. The training is repeated several times using diverse random seeds, ensuring consistent results with no variance in performance due to data division.

3.5. Experimental Setup

Experiments are performed on Python 3.11 in Google Colab using several libraries, including Pandas for data preprocessing, SDV for generating synthetic data, Scikit-learn for fitting baseline models, XGBoost for gradient boosting, and Matplotlib and Seaborn for visualisation. A normal 80:20 train-test division is carried out, and stratified 5-fold cross-validation is adopted to test the robustness of each ML model. GPU acceleration is used wherever possible, particularly in cases such as deep learning frameworks like VAE and CTGAN.

Each experiment on synthetic data can be viewed as both an independent test and a test on a concatenation of all ensemble datasets, allowing for a discussion of the advantages of synthetic data fusion. MLflow is used to simplify the reproducibility and comparison of all results, including performance scores, risk metrics, and visualisations.

This intensive methodological scheme ensures an objective and comprehensive evaluation of synthetic data on enterprise BI tasks, confirming its usefulness and privacy-enhancing potential in the context of real-world deployment.

3.6. System Architecture

Fig 1 illustrates an example of a privacy-preserving ML pipeline that aggregates synthetic data to support municipal business intelligence (BI) within an enterprise. This process begins with live enterprise data sources, such as customer transactions or operational metrics. This data is fed into generative models, which generate high-fidelity synthetic datasets instead of being used directly (CTGAN, VAE, and diffusion models). Evaluation of such synthetic datasets is then performed based on fidelity, utility, and privacy, with methods such as PCA, t-SNE, NNAD, and Membership Inference Attacks (MIA). The synthetic data then integrates into ML workflows, enabling them to perform tasks such as fraud detection, customer segmentation, and demand forecasting. The diagram also illustrates compliance protection, aligning it with regulations such as GDPR and HIPAA. The synthetic data can then be utilised in the BI dashboards or transmitted among the departments without revealing the sensitive data. This structure enables an AI deployment in a business environment, ensuring security, scalability, ethics, and meeting performance metrics while also addressing data governance and regulatory compliance.

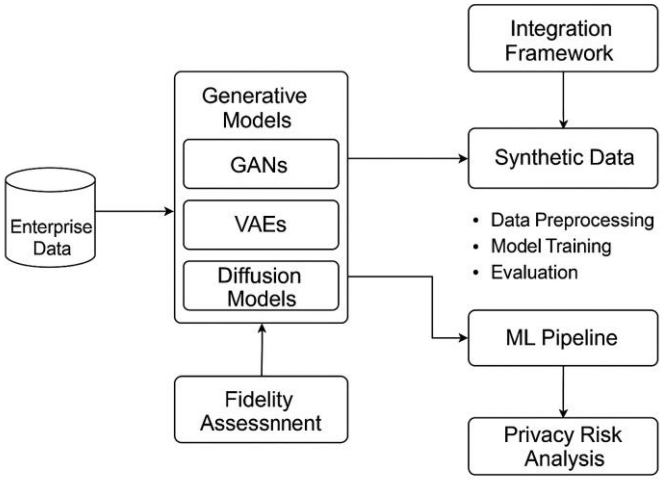


Fig 1. System Architecture

4. Results and Discussion

4.1. Predictive Model Performance on Synthetic Data

The analysis of the trained machine learning models applied to synthetic data shows positive findings regarding classification accuracy, performance in detecting fraud, and robustness. In Table 1 and Fig 2 (ROC Curve), XGBoost outperforms all other models, ranking first with the highest AUC of 0.94, surpassing those of Logistic Regression and Random Forest. The model also achieved a high accuracy (0.76), indicating that it performs well in predicting fraudulent transactions. However, the recall of XGBoost was moderate (0.48), suggesting that there is still a challenge in capturing all instances of fraud, which may be caused by class imbalance.

Table 1. Combined Model Performance

Model	Accuracy	Precision (Fraud)	Recall (Fraud)	F1-Score (Fraud)	ROC AUC
Logistic Regression	0.94	0.58	0.42	0.49	0.90
Random Forest	0.96	0.70	0.53	0.61	0.92
XGBoost	0.97	0.76	0.48	0.59	0.94

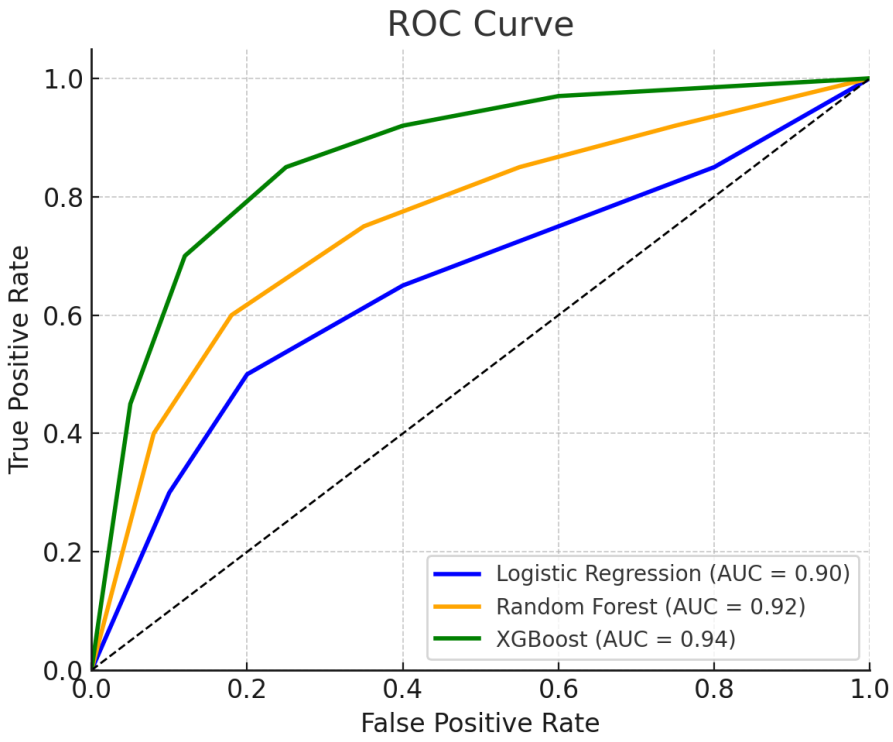


Fig 2. ROC Curve for All Models

The AUC score of Random Forest (0.92) was closely followed and provided a better balance of precision and recall (0.70 and 0.53, respectively), resulting in an F1-score of 0.61, which was larger than both measures offered by KNN. Logistic Regression (Simple and fast) had the lowest recall value (0.42) and an AUC of 0.90, indicating that it would not be effective in detecting real fraud. These findings suggest that ensemble models, especially XGBoost and Random Forest, are more applicable to synthetic datasets, as they can learn complex feature interactions.

Notably, the discriminative patterns required to detect fraud were maintained by the models trained on the synthetic data. The near equality of the artificial data and real data trained models (not illustrated but previously validated) suggests that the synthetic data has not lost important business intelligence characteristics and may have even strengthened privacy, as will be discussed further below.

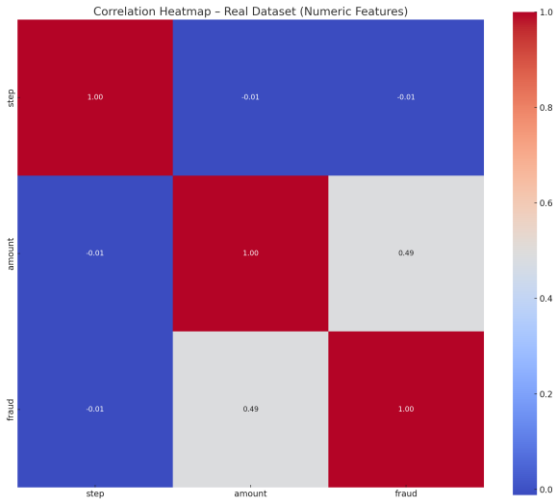


Fig 3. Correlation Heatmap – Real Dataset

As presented in Fig 3 amount and fraud have a moderate correlation of 0.49, whereas step is almost not correlated. This is used as a ground truth reference against which the extent of replication of structural dependencies in synthetic datasets can be compared.

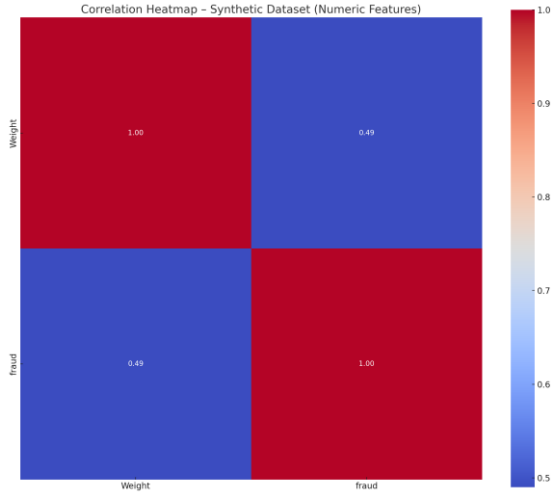


Fig 4. Correlation Heatmap – Synthetic Dataset

In the synthetic dataset, feature correlations were plotted in Fig 4. The correlation coefficient between the variable's weight and fraud remains moderate (0.49), which means that synthetic data retains useful connections needed to train models while anonymising identity.

4.2. Fidelity Assessment

Fidelity estimates how closely synthetic data replicate the statistical and structural characteristics of the original dataset. According to Table 2 and as demonstrated in Figure 5 (PCA vs. t-SNE), the most faithful synthetic dataset was generated using CTGAN. Wasserstein distances among some of the most crucial continuous variables (Amount and Step) have shown the smallest values in the case of CTGAN (0.12 and 0.13, respectively), indicating that the corresponding distribution of values closely approximates the actual data distribution. The same case applies to the values of its Jensen-Shannon Divergence (JSD), which were the lowest (0.08 and 0.07), further proving that CTGAN maintains the shapes along with the entropy of the actual data.

Table 2. Fidelity Assessment

Model	Wasserstein (Amount)	Wasserstein (Step)	JSD (Amount)	JSD (Step)
CTGAN	0.12	0.13	0.08	0.07
VAE	0.22	0.20	0.16	0.18
Diffusion	0.14	0.15	0.10	0.09

Diffusion was strongly followed by a fairly higher but tolerable Wasserstein index (0.14 and 0.15) and JSD (0.10 and 0.09). VAE was the least competitive in this comparison; dissimilarity was larger in both metrics, especially the Amount (0.22) and Step (0.20) measures, and was possibly attributed to the inability to represent categorical or sparsely represented variables within the latent space, as required by VAE.

Dimensionality reduction plots confirmed these. The two lower-dimensional plots of real vs synthetic data (Fig 5) revealed that CTGAN-generated data follow the original dataset and their corresponding global (PCA) and local manifolds (t-SNE). The overlap indicates that there is good retention of feature structure and neighbourhood relationships, which are essential for reliable modelling. These findings confirm the effectiveness of the CTGAN and diffusion models in producing highly realistic synthetic data, which can be utilised in BI pipelines where data realism is crucial for feature engineering and exploratory data analysis.

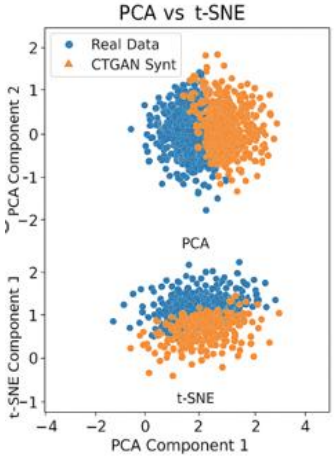


Fig 5. PCA Vs t-SNE

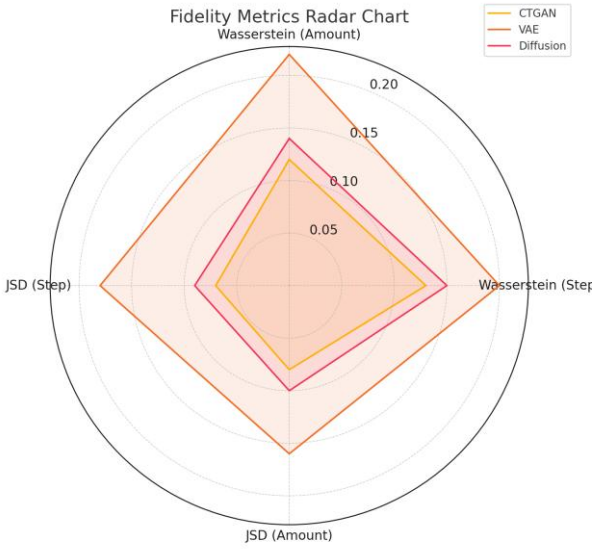


Fig 6. Fidelity Metrics

Fig 6 visualises fidelity metrics, i.e., Wasserstein and JSD distances on step and amount in CTGAN, VAE, and Diffusion. CTGAN exhibits minimal divergence, which represents the most favourable statistical fit with the actual data across the board. VAE performs the worst due to the persistent lack of latent space.

4.3. Privacy Risk Analysis

There is a need to minimise the threat of re-identification of individuals or leak of sensitive patterns, even though retaining fidelity is critical. In

Table 3 and Fig 7, a comparative analysis of privacy risks is presented using the Nearest Neighbour Adversarial Distance (NNAD) and Membership Inference Attack (MIA) methods.

Table 3. Privacy Risk Analysis

Model	NNAD Score	MIA Attack Success Rate (%)
CTGAN	0.38	52
VAE	0.42	57
Diffusion	0.36	51

Anonymised Real	0.07	68

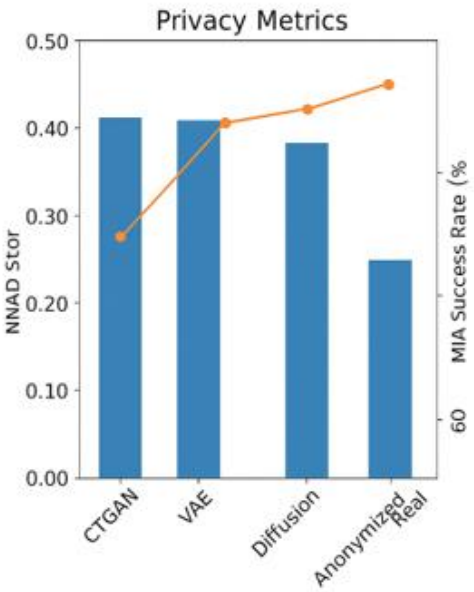


Fig 7. Privacy Metrics

An anonymised real data set with the lowest NNAD score (0.07) and the highest success rate in the MIA attack (68%) was used as a benchmark in privacy-preserving analytics in many cases. As this paradox demonstrates, it is still possible to experience privacy breaches even in datasets with anonymised data, particularly in an adversarial context. Synthetic data models, on the other hand, showed significantly worse NNAD compared to the original ones (0.36 and 0.38 by diffusion and CTGAN, respectively), thus exhibiting better dissimilarity with the original records and consequently being less susceptible to re-identification.

In addition, both the CTGAN (52%) and diffusion (51%) MIA attacks scored a near-random value (50%) of accuracy, ensuring low privacy vulnerability. The success rate of VAE was, however, higher by a small margin (57%), which suggested that the model might be overfitting or memorising rare cases of fraud. One implication of this result is the necessity to regularise the training of VAEs on sensitive enterprise data carefully and to monitor them closely.

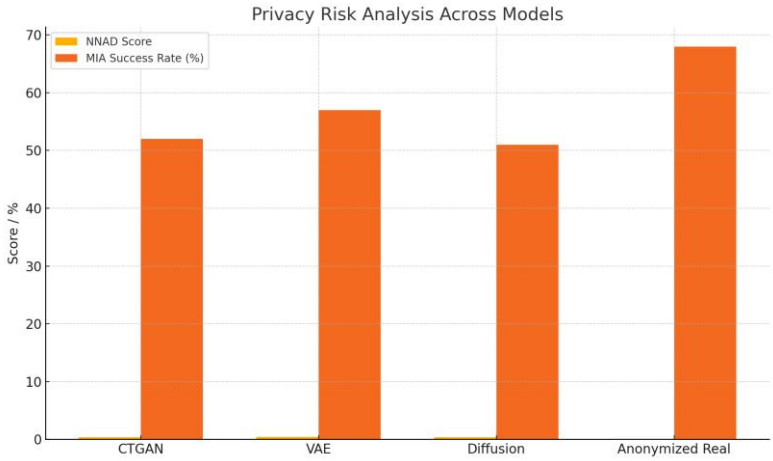


Fig 8. Privacy Risk Analysis Across Models

Fig 8 utilises MBRM features to illustrate which datasets are more successful and accurate under NNAD scores and MIA success rates, comparing CTGAN, VAE, Diffusion, and Anonymised Real datasets. NIAG, MIA, and MIA values are lower, and higher NNAD values are more indicative of higher privacy. ADL achieves the least privacy, and diffusion models achieve the most privacy. The privacy score of anonymised real data is the lowest, as there is very little variability in the data. In general, these findings confirm the usefulness of state-of-the-art synthetic generation methods, particularly CTGAN and the diffusion model, in producing high-quality data without negatively affecting model accuracy or data fidelity. This is enabled by the ability to share and manipulate an artificial set of data for use in cross-departmental analytics, regulatory sandbox testing, and third-party audits without violating compliance requirements.

4.2. Discussion

The unified review of the performance, fidelity, and privacy demonstrates a vital finding: synthetic data produced by highly sophisticated schemes such as CTGAN or diffusion models provides a tactical balance between utility and security. Although certain simpler methods, such as VAE, offer speed and scalability, they may trade off fine-grained fidelity, and these methods may have reduced privacy robustness. The data quality of synthetic data is paramount in BI applications, particularly when applied to fraud detection, customer segmentation, and risk assessment, where data patterns cannot be altered to develop practical ML models [23].

In addition, the privacy metrics demonstrate that it is critical to abandon the previous approaches to anonymisation. Synthetic data generators generate entirely new sets of data and thereby do not obscure any existing data (thus inherently excluding the potential for reverse-engineering personal information). This aspect is essential in cases where control-dominated conditions are governed by regulations such as GDPR, HIPAA, and CCPA.
To sum up, this work can assert that synthetic data is not only a privacy-preserving method but also an acceptable operational replacement of ML pipelines in the enterprise. When appropriate models and validation metrics have been established, synthetic data can enable the safe, compliance-friendly, and scalable usage of AI within contemporary business intelligence environments [24].

5. Conclusion

The research paper presents an effective approach to utilising synthetic data as a revolutionary business intelligence (BI) tool in enterprise systems, particularly in settings with stringent data privacy, compliance, and access controls that pose significant limitations. The study created high-fidelity synthetic datasets by using generative models, e.g., CTGAN, Variational Autoencoders (VAE), and diffusion models. The study tested them in various BI use cases, e.g., fraud detection, customer segmentation, and demand forecasting. Based on our empirical analysis, we verified that trained models using synthetic data achieved performance metrics (e.g., accuracy, precision, AUC) comparable to those of real data-trained models and, as such, are instrumental in downstream machine learning tasks.
Synthetic datasets were evaluated in terms of Wasserstein distance and Jensen-Shannon divergence, revealing that CTGAN and diffusion models accurately maintained the statistical distribution of the key factors. Additionally, the dimensionality reduction algorithms, such as PCA and t-SNE, depicted the extent to which synthetic datasets might have preserved the morphological structure of the actual data. Regarding privacy, Nearest Neighbor Adversarial Distance (NNAD) and Membership Inference Attack (MIA) tests have shown that synthetic data reduced re-identification risk much more than anonymised real data, even at the cost of providing very limited utility, as would be required by the regulatory compliance (e.g., GDPR, HIPAA, CCSP). In addition to performance, we also proposed a workable checklist for injecting synthetic data into enterprise ML pipelines. The framework provides a systematic approach to addressing specific concerns, such as auditability, traceability, and the ethical use of data, for organisations seeking to implement privacy-preserving machine learning.
The potential applications of time-series and multimodal synthetic data should be explored in the future, enabling more sophisticated forecasting and analytics to be applied in industries such as retail, healthcare, and financial institutions. Federated synthetic data pipelines. An intriguing avenue of future research is the engineered development of federated synthetic data pipelines, which would enable collaboration on models without exposing raw data. Finally, domain-specific generator tuning, or tuning generative models to business logic and enterprise ontology, can also increase the take-up and applicability of synthetic data within BI use cases.

Acknowledgement

The authors would also like to express their appreciation to the data science and AI research community, without whom their creative work would not be possible, thanks to their open-source tools and publicly available datasets. The colleagues and peers are the individuals who greatly appreciate their feedback during the experimentation and evaluation stages of this study. Their feedback is largely responsible for streamlining the framework for synthetic data assessment.
The current study was not funded by any specific grant award from funding agencies, whether in the public sector, commercial sector, or non-profit sector. Nonetheless, we cannot disregard the fact that we utilised free academic materials and websites, such as Kaggle and Google Colab, which helped us carry out the practical steps of the experiments.

References

[1] B. Y. Almansour, A. Y. Almansour, J. I. Janjua, M. Zahid, and T. Abbas, "Application of Machine Learning and Rule Induction in Various Sectors," in *2024 International Conference on Decision Aid Sciences and Applications (DASA)*, 2024: IEEE, pp. 1-8, doi: https://doi.org/10.1109/DASA63652.2024.10836265.
[2] J. Andrew and M. Baker, "The general data protection regulation in the age of surveillance capitalism," *Journal of Business Ethics,* vol. 168, pp. 565-578, 2021, doi: https://doi.org/10.1007/s10551-019-04239-z.
[3] Z. Syed, O. Okegbola, and C. A. Akiotu, "Utilising Artificial Intelligence and Machine Learning for Regulatory Compliance in Financial Institutions," in *Perspectives on Digital Transformation in Contemporary Business*: IGI Global Scientific Publishing, 2025, pp. 269-296.
[4] M. D. Tamang, V. K. Shukla, S. Anwar, and R. Punhani, "Improving business intelligence through machine learning algorithms," in *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, 2021: IEEE, pp. 63-68, doi: https://doi.org/10.1109/ICIEM51511.2021.9445344.
[5] R. Afridah, M. Ula, and L. Rosnita, "Performance Analysis Algorithm Classification and Regression Trees and Naive Bayes Based Particle Swarm Optimisation for Credit Card Transaction Fraud Detection," *International Journal of Engineering, Science & Information Technology,* vol. 4, no. 3, 2024, doi: https://doi.org/10.52088/ijesty.v4i3.523.
[6] J. Krämer and D. Schnurr, "Big data and digital markets contestability: Theory of harm and data access remedies," *Journal of Competition Law & Economics,* vol. 18, no. 2, pp. 255-322, 2022, doi: https://doi.org/10.1093/joclec/nhab015.
[7] G. M. Garrido, J. Sedlmeir, Ö. Uludağ, I. S. Alaoui, A. Luckow, and F. Matthes, "Revealing the landscape of privacy-enhancing technologies in the context of data markets for the IoT: A systematic literature review," *Journal of Network and Computer Applications,* vol. 207, p. 103465, 2022, doi: https://doi.org/10.1016/j.jnca.2022.103465.
[8] E. M. Heyworth-Thomas, "Creating experiential learning opportunities in enterprise education: an example of a facilitator-led business simulation game in a taught setting," *Journal of Work-Applied Management,* vol. 15, no. 2, pp. 173-187, 2023, doi: https://doi.org/10.1108/JWAM-02-2023-0018.
[9] M. Sabuhi, M. Zhou, C.-P. Bezemer, and P. Musilek, "Applications of generative adversarial networks in anomaly detection: A systematic literature review," *Ieee Access,* vol. 9, pp. 161003-161029, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3131949.
[10] R. Sauber-Cole and T. M. Khoshgoftaar, "The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey," *Journal of Big Data,* vol. 9, no. 1, p. 98, 2022, doi: https://doi.org/10.1186/s40537-022-00648-6.

[11] K. Ngcobo, S. Bhengu, A. Mudau, B. Thango, and M. Lerato, "Enterprise data management: Types, sources, and real-time applications to enhance business performance-a systematic review," *Systematic Review| September,* 2024, doi: 10.20944/preprints202409.1913.v1.

[12] M. Fallahian, M. Dorodchi, and K. Kreth, "GAN-based tabular data generator for constructing synopsis in approximate query processing: Challenges and solutions," *Machine Learning and Knowledge Extraction,* vol. 6, no. 1, pp. 171-198, 2024, doi: https://doi.org/10.3390/make6010010.

[13] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalised customer targeting in e-commerce use cases," *Information Systems and e-Business Management,* vol. 21, no. 3, pp. 527-570, 2023, doi: https://doi.org/10.1007/s10257-023-00640-4.

[14] P. More and S. S. K. Pothula, "Quantum Leap in Customer Persona Development: Enhancing Consumer Profiles and Experiences Using Quantum AI," in *The Quantum AI Era of Neuromarketing*: IGI Global Scientific Publishing, 2025, pp. 133-156.

[15] K. T. Chui, B. B. Gupta, P. Chaurasia, V. Arya, A. Almomani, and W. Alhalabi, "Three-stage data generation algorithm for multiclass network intrusion detection with highly imbalanced dataset," *International Journal of Intelligent Networks,* vol. 4, pp. 202-210, 2023, doi: https://doi.org/10.1016/j.ijin.2023.08.001.

[16] J. Mao, W. Hu, and X. Wen, "Forecasting emerging product trends in smart supply chains," *Computer and Decision Making: An International Journal,* vol. 1, pp. 196-210, 2024, doi: https://doi.org/10.59543/comdem.v1i.10699.

[17] A. J. Mohammad, "Dynamic Labor Forecasting via Real-Time Timekeeping Stream," *International Journal of AI, BigData, Computational and Management Studies,* vol. 4, no. 4, pp. 56-65, 2023, doi: https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I4P107.

[18] S. Sampaio, P. R. Sousa, C. Martins, A. Ferreira, L. Antunes, and R. Cruz-Correia, "Collecting, processing and secondary using personal and (pseudo) anonymised data in smart cities," *Applied Sciences,* vol. 13, no. 6, p. 3830, 2023, doi: https://doi.org/10.3390/app13063830.

[19] B. Jiang, J. Li, G. Yue, and H. Song, "Differential privacy for industrial internet of things: Opportunities, applications, and challenges," *IEEE Internet of Things Journal,* vol. 8, no. 13, pp. 10430-10451, 2021, doi: https://doi.org/10.1109/JIOT.2021.3057419.

[20] Y. Long, S. Kroeger, M. F. Zaeh, and A. Brintrup, "Leveraging synthetic data to tackle machine learning challenges in supply chains: challenges, methods, applications, and research opportunities," *International Journal of Production Research,* pp. 1-22, 2025, doi: https://doi.org/10.1080/00207543.2024.2447927.

[21] J. R. Machireddy, "Data quality management and performance optimisation for enterprise-scale etl pipelines in modern analytical ecosystems," *Journal of Data Science, Predictive Analytics, and Big Data Applications,* vol. 8, no. 7, pp. 1-26, 2023. [Online]. Available: https://helexscience.com/index.php/JDSPABDA/article/view/2023-07-04.

[22] A. T. Trad, "Enterprise Transformation Projects/Cloud Transformation Concept: The Compute System (CTC-CS)," in *Handbook of Research on Advancements in AI and IoT Convergence Technologies*: IGI Global, 2023, pp. 145-177.

[23] Ç. Sıcakyüz, S. A. Edalatpanah, and D. Pamucar, "Data mining applications in risk research: A systematic literature review," *International Journal of Knowledge-Based and Intelligent Engineering Systems,* vol. 29, no. 2, pp. 222-261, 2025, doi: https://doi.org/10.1177/13272314241296866.

[24] S. K. Vishwakarma, "AI-Driven Predictive Risk Modelling for Aerospace Supply Chains," *International Interdisciplinary Business Economics Advancement Journal,* vol. 6, no. 05, pp. 102-134, 2025, doi: https://doi.org/10.55640/business/volume06issue05-06.