GAN-Based Synthetic Data Augmentation and SVM Driven Customer Segmentation in Retail BI Systems

N. Janakibai[1], M.SrinivasuRao Naidu[2], Y. Jagadesh Kumar[3]
[1,2] *Department of Computer Science and Engineering (CSE) Sri sivani institute of technology, Srikakulam. Andhra Pradesh, INDIA.*
[3] *Head of The Department, Department of Computer Science and Engineering (CSE) Sri sivani institute of technology, Srikakulam. Andhra Pradesh, INDIA.*

***Abstract*—The integration of synthetic data generation and machine learning classification offers a transformer approach to retail analytics and decision intelligence. This study Proposes a framework that leverages Generative Adversarial Networks (GANs) to-produce realistic synthetic retail transaction datasets, addressing challenges of limited historical data and privacy concerns. The generated data is subsequently classified using Support Vector Machines (SVM), enabling segmentation of customers, prediction of-product demand categories, and identification of purchasing patterns. These classified insights are embedded into interactive Business Intelligence (BI) dashboards, providing stakeholders with actionable visualizations such as sales heat maps, customer segmentation cards, and trend forecasts. The proposed system demonstrates how Gan driven synthetic data, combined with SVM classification, can enhance scalability,preserve confidentiality, and deliver predictive intelligence for retail decision-making. By bridging advanced data science techniques with BI visualization, this framework empowers organizations to simulate diverse market scenarios, anticipate consumer behavior, and optimize strategic planning in data-constrained environments.The rapid expansion of AI and ML applications has heightened the demand for large, diverse, and high-quality datasets. The purpose of the Synthetic Data volume generator is overcoming the data scarcity or protecting sensitive information. The methods that are used in synthetic data volume generators are Generative Adversarial network(GAN) and Variational Auto encoders. A Variational Auto encoder is a machine learning model that can generate new data based on the data its trained on. Variational Auto encoders are used in many applications like image generation, text generation and data denoising.**

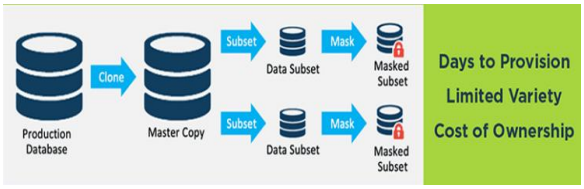
***Index Terms*—GAN, SVM and BI-Dashboard**

I. INTRODUCTION

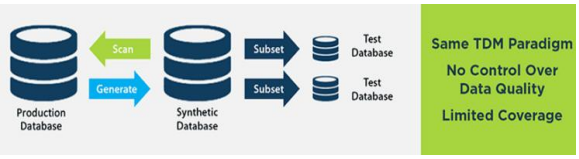
The Synthetic Data Volume Generator project aims to design and develop a cutting-edge tool for generating large volumes of synthetic data, addressing the growing need for high-quality, realistic, and cost-effective data for applications like data analytics, machine learning, and software testing. The project's primary objectives include developing a scalable synthetic data generation engine, creating a user-friendly interface, integrating with various data platforms, and ensuring generated data meets quality standards. By completing this project, we expect to deliver a highly scalable and efficient synthetic data generation engine, a user-friendly interface, improved data quality, and enhanced support for data analytics and machine learning applications. This project will provide a robust solution for organizations to generate synthetic data that mimics real-world data patterns, allowing them to test, train, and validate their applications without the need for actual sensitive data. The Synthetic Data Volume Generator will offer a wide range of benefits, including reduced data management costs, improved data privacy and security, and accelerated application development and testing. Furthermore, the project will ensure that the generated synthetic data is highly realistic, diverse, and representative of real-world data, making it an ideal solution for various industries, including finance, healthcare, retail, and more. Overall, the Synthetic Data Volume Generator project has the potential to revolutionize the way organizations generate and utilize synthetic data, enabling them to make better decisions, improve operational efficiency, and drive business growth. The rapid expansion of AI and ML

applications has heightened the demand for large , diverse ,and high quality datasets. The purpose of the Synthetic Data volume generator is overcoming the data scarcity or protecting sensitive information. The methods that are used in synthetic data volume generators are Generative Adversarial network (GAN) and Variational Auto encoders. A Variational Auto-encoder is a machine learning model that can generate new data based on the data its trained on. Variational Auto encoders are used in many applications like image generation, text generation and data denoising.

Artificial Data: Synthetic data is not real data but is created arithmetically or through simulations to resemble real- world data. Mimicking Real Data: It aims to capture the statistical properties and patterns of real data, allowing for analysis and training of models without using actual, sensitive information. Methods of Synthetic Data Generation : Synthetic data generation is much faster than manual data creation and can produce higher data volumes for load and performance testing. It's an essential technology for reducing test cycle time and implementing shift-left testing strategies. However, not all synthetic data generation methods and technologies are the same. It's important to understand the differences and their impact on test data quality. Synthetic data generation systems operate in two fundamentally different ways: 1. A synthetic data replica is produced by scanning and profiling a production data source 2. Synthetic data is designed and generated dynamically to meet test case requirements The first method is often associated with the use of deep learning to model and profile a production database to produce a statistically representative synthetic data replica. This results in a secure and private synthetic version of production data without its sensitive information. Then it can safely be used for data analytics and business intelligence use cases. However, statistically representative synthetic data is not very suitable for software testing and quality assurance. The reason for this is simple: A synthetic replica has the same limitations as the original production database from which it was derived. If data patterns, permutations and variations are missing from the production database, they will also be missing from the synthetic replica. Using a synthetic replica of production data for testing is a modern variation of the traditional Test Data Management (TDM) paradigm. When using a conventional TDM system, the QA team must copy, subset, mask, reserve, and refresh production data before it can be used for automated testing. In addition to the protracted provisioning time imposed by TDM systems, missing data variations reduce the overall test data quality and ultimately limit test coverage. Using a synthetic data replica for software testing is just a faster and more automated version of the same tdm approach. The provisioning steps are streamlined, but the result is the same – missing data variations reduce test data quality and coverage.



Production Data for Testing



Synthetic Data for Testing

II. LITERATURE SURVEY

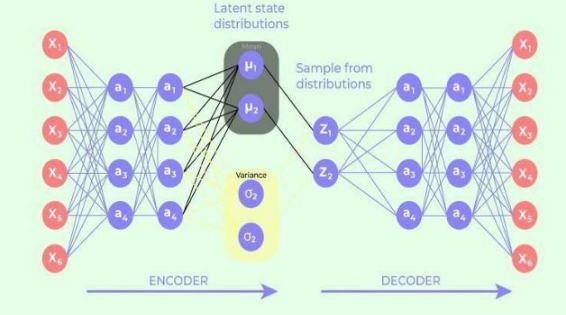
Synthetic data consists of artificially generated data. When data are scarce, or of poor quality, synthetic data can be used, for example, to improve the performance of machine learning models. Generative adversarial networks (GANs) are a state-of-the-art deep generative models that can generate novel synthetic samples that follow the underlying data distribution of the original dataset. Reviews on synthetic data generation and on GANs have already been written. However, none in the relevant literature, to the best of our knowledge, has explicitly combined these two topics. This survey aims to fill this gap and provide useful material to new researchers in this field. That is, we aim to provide a survey that combines synthetic data generation and GANs, and that can act as a good and strong starting point for new researchers in the field, so that they have a general overview of the key contributions and useful references. We have conducted a review of the state-of-the-art by querying four major databases: Web of Sciences (WoS), Scopus, IEEE Xplore, and ACM Digital Library. This allowed us to gain insights into the most relevant

authors, the most relevant scientific journals in the area, the most cited papers, the most significant research areas, the most important institutions, and the most relevant GAN architectures. GANs were thoroughly reviewed, as well as their most common training problems, their most important breakthroughs, and a focus on GAN architectures for tabular data. Further, the main algorithms for generating synthetic data, their applications and our thoughts on these methods are also expressed. Finally, we reviewed the main techniques for evaluating the quality of synthetic data (especially tabular data) and provided a schematic overview of the information presented in this paper. Data are ubiquitous and can be a source of great value. However, to create such value, the data needs to be of high quality. In addition, when dealing with sensitive data (e.g., medical records or credit datasets), the privacy of the data must be ensured without sacrificing quality. enabling industries to address data privacy, scalability, and cost-effectiveness challenges, particularly in AI and machine learning applications. Synthetic data, generated to mimic real-world data, is used across various industries to train AI models, test software, and conduct research while protecting sensitive information and addressing data scarcity.These tools generate artificial data that replicates the statistical properties and characteristics of real-world data, without containing actual, identifiable information.

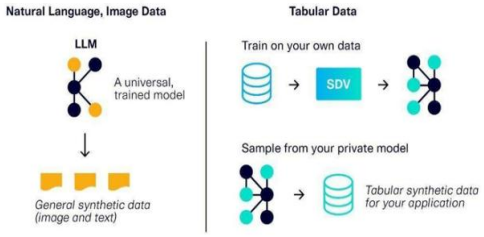
III. PROPOSED ARCHITECTURE

Algorthims Used for Data Generation Generative Adversarial Networks Generative adversarial networks (GANs)are a framework that uses an adversarial process to estimate generative deep learning models, proposed by Ian J. Goodfellow et al. in 2014. These structures have been adapted and improved over the last years and are now very powerful. GANs are currently capable of painting, writing, composing, and playing, as we will see in this section. Therefore, GANs are first analyzed in more detail to reveal how they work. In the main training difficulties and their solutions are examined. Finally, in, the main GAN architectures are shown to demonstrate their capabilities, while in GANs for tabular data are presented. A GAN is constituted by two models: a generator model G that tries to generate samples that follow the underlying distribution of the data. Nonetheless, these observations are suitably different from the ones in the dataset (i.e., they should not simply reproduce observations that already occur in the dataset). There is also a discriminator model D that, given an observation (from the original dataset or synthesized by the generator), classifies it as fake (produced by the generator) (Typically, the models used for the generator and discriminator are neural networks. As such, we normally refer to G and D as networks. However, in theory, the models need not be a neural network. Indeed, in , the term “model” is used. Nonetheless, they note that the “adversarial modeling framework is most straightforward to apply when the models are both multi-layer perceptrons”) or real. An important thing to consider is that G and D compete against each other. While G generates similar data points to those in the original data, with the aim of deceiving the discriminator, D attempts to distinguish the generated from the real observations. To describe in more detail how the networks are trained, the training was split into the training of the discriminator and of the generator separately. Training the discriminator consists of creating a dataset with instances generated by G and data points from the original dataset. The discriminator outputs a probability (continuous value between 0 and 1) that indicates whether a given observation came from the original data (0 means that the discriminator is 100%100% certain that the given example was synthesized, while 1 means the exact opposite). The training of the generator is more complicated. G is given as the input random noise (The term latent space is typically used to designate G’s input space.), commonly from a multivariate normal distribution, and the output is a data point with the same features of the original dataset. Deep Conventional Generative Adversarial Network, DCGAN, is a GAN architecture that combines conventional layers (A conventional layer is a layer that uses a convolution operation. A convolution, in terms of computer vision tasks, consists of a filter (represented by a matrix) that slides through the image pixels (also represented by a matrix) and performs matrix multiplication. This is useful in computer vision tasks because applying different filters to an image (by means of a convolution) can help, for example, detect edges, blur the image, or even remove noise), which are commonly used in computer vision tasks, with GANs. Redford et al., in , have brought together the success

of Conventional Neural Networks (CNN) in supervised learning tasks with the then emerging GANs. Nowadays, the use of conventional layers in GANs for image generation is quite common, but at that time, 2016, this was not the case. Therefore, the use of conventional layers in the GAN structure is still a powerful tool for handling image data. Information Maximizing Generative Adversarial Network, Info GAN, is a GAN extension proposed by Chen et al. in , that attempts to learn disentangled information. That is, to give semantic meaning to features in the latent space . Info GAN can successfully recognize writing styles from handwritten digits in the MNIST dataset, detect hairstyles or eyeglasses in the Celeb A dataset, or even background digits from the central digit in the SVHN dataset. Coupled Generative Adversarial Networks, Co GAN, proposed in by Ming -Yue Lin and Once l Tuzel, use a pair of GANs instead of only one GAN. The Co GAN was used to learn the joint distribution of multi-domain images, which was achieved by the weight-sharing constraint between the two GANs. In addition, sharing weights requires fewer parameters than two individual GANs, which, in turn, results in less memory consumption, less computational power, and fewer resources. Auxiliary Classifier Generative Adversarial Network, AC-GAN , is a GAN extension proposed by Odena et al. that modifies the generator to be class dependent (it takes class labels into account) and adds an auxiliary model to the discriminator whose purpose is to reconstruct the class label. The results in show that such an architecture can generate globally coherent samples that are comparable, in diversity, to those of the ImageNet dataset . Stacked Generative Adversarial Network, StackGAN, proposed in by Zhang et al., is another extension of GANs that can generate images from text descriptions. This generation of photo realistic images is decomposed into two parts. First, the STAGE-I GAN sketches a primitive shape and colors based on the input text. Next, the Stage-II GAN uses the same text description as the STAGE-I GAN and its output as input and generates high-resolution images by refining the output images by STAGE-I GAN. Their work has led to significant improvements in image generation. Hence, the training sample size was set to 2000 on the lower end in the initial iteration which was 4% of the maximum size and then increased to 5000 (10%), 10000 (20%), 20000 (40%) and 50000 (100%) for every dataset. Over the past year, we've worked closely with a select set of customers to get more insight into how enterprises operate, and understand the best ways to incorporate synthetic data. In 2023 alone, we made 62 software releases across our libraries, and developed a framework for SDV Enterprise.



Variational Auto Encoder



IV. RESULTS

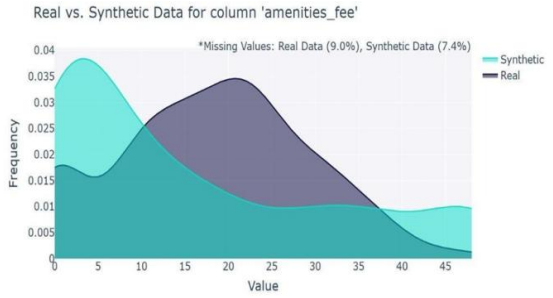
The outcomes presented in this section are the summary of our results after narrowing down on the SDGs that are generating datasets that are close to real data based on their evaluation metrics. Table 2 tabulates the comparison of minority class percentages in the original dataset and the datasets generated using different generators

Tabular Format of SDV

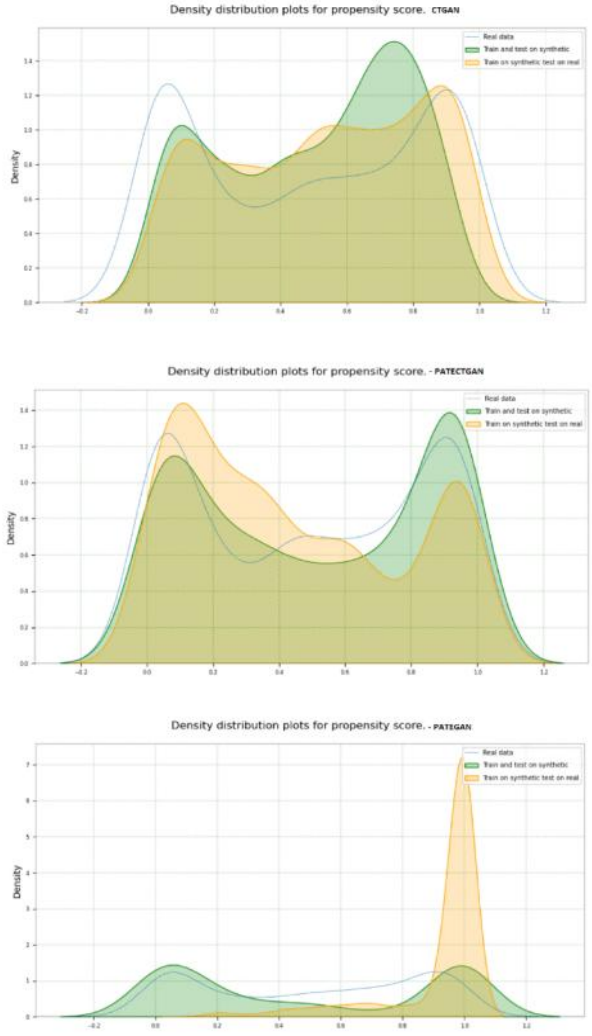
	guest_email	has_rewards	room_type	amenities_fee	\
0	michaelsanders@shaw.net	False	BASIC	37.89	
1	randy49@brown.biz	False	BASIC	24.37	
2	webermelissa@neal.com	True	DELUXE	0.00	
3	gsims@terry.com	False	BASIC	NaN	
4	misty33@smith.biz	False	BASIC	16.45	
..	...	...	...	...	
495	laurabennett@jones-duncan.net	False	BASIC	8.71	
496	johnny71@cook.info	False	BASIC	16.31	
497	ygarcia@ballard-lopez.net	False	BASIC	30.59	
498	thomasdale@hall.com	False	BASIC	1.93	
499	danieltaylor@harper.com	False	BASIC	3.84	

	checkin_date	checkout_date	room_rate	\
0	27 Dec 2020	29 Dec 2020	131.23	
1	30 Dec 2020	02 Jan 2021	114.43	
2	17 Sep 2020	18 Sep 2020	368.33	
3	28 Dec 2020	31 Dec 2020	115.61	
4	05 Apr 2020	NaN	122.41	
..	...	...	...	
495	04 Jan 2021	06 Jan 2021	103.25	
496	24 Aug 2020	26 Aug 2020	115.81	
497	11 Nov 2020	13 Nov 2020	141.61	

Result of Prod data

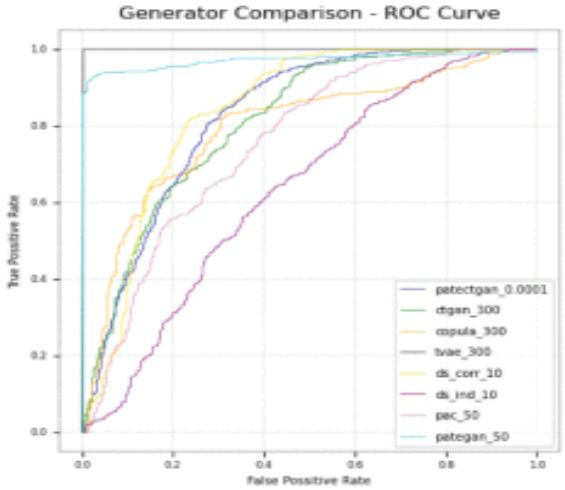


Synthetic data generation by oversampling the minority class in input sample.

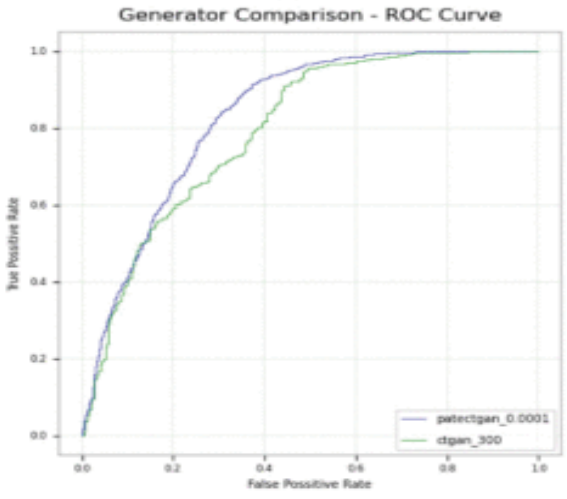


SDV Enterprise is built for scalability. Enhancements to the modeling and sampling engines allow enterprise users to model much more complex schemas, potentially with hundreds of interconnected tables. Our team has thoroughly tested the engines on common patterns in database design, including bridge tables, star schemas, and snowflake schemas. Our early SDV Enterprise customers have also shared their schema designs with us, in which common elements of database design have become elements of a larger application with even more inter connectedness! SDV Enterprise can handle these too. SDV Enterprise deeply understands data. For human-created concepts and industry-specific terms, SDV Enterprise offers a boost in data quality. It is designed to parse and extract the deeper meaning behind the data, enabling it to create realistic synthetic data that preserves this meaning Our team continues to add support for data across domains, from the automotive industry, to finance, to healthcare.

Comparison of propensity distribution of ML model for real data (blue), synthetic data (green) and model trained on synthetic data and tested on real data for datasets generated using CTGAN, PATECTGAN and PATEGAN generators.



(a)



(b)

V. CONCLUSION

From our experiments we observe that no single SDG can handle all input sample scenarios perfectly. Therefore, our findings cannot be generalized. Based on results. with the smallest noise multiplier of 0.0001 and € = 2.5, yields better results for the classification model, a noise_multiplier of 0.0001 with € = 5 is slightly better in terms of utility. It is important to note that increasing the noise multiplier for better privacy comes at the cost of data utility and introduces more bias into the dataset. CTGAN exhibits similar behavior. The commercial generator provided by Gretel AI exhibited no variation. Consequently, the selection of privacy-enhancing parameters must be contingent on the requirements of the business.

REFERENCE

[1] hari, A., & Soraya, M. (2010). Workload generation for Multimedia Tools and Applications,
[2] Computers & Industrial Engineering, 19(1–4), 582–586. HTTP://dx.doi.org/10.1016/0360-8352(90)90185-OADL (Advanced Distributed Learning). (2015). x API specification.