International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 05 | May 2026 www.irjet.net p-ISSN: 2395-0072

Synthetic Data Generation and Its Role in Machine Learning Model Development

Shreyas N[1], Supriya S G[2], Sneha P URS[3], Suhas Kandalam[4], Dr.Ramesh Sekaran[5]

Dept of CSE, School of Engineering, Dayananda Sagar University, Karnataka, India

---

Abstract- *Training data is a fundamental research limit to machine learning by constraining its performance in terms of its availability, quality, and representativeness. Real-world data is often small, skewed in representation, sensitive, or expensive to label (and this list is not exhaustive) and this hinders the extrapolability of the state-of-the-art models. This paper focuses on synthetic data generation as a methodological approach to such challenges. We review existing methods such as the Synthetic Minority Over-sampling Technique (SMOTE), Generative Adversarial Networks (GANs), Generative Adversarial Conditioned Tabular GAN (CTGAN), Variational Auto encoders (VAEs), and diffusion-based generative models. We discover three essential research gaps: We have not yet found generation methods capable of faithfully encoding intricate data distributions, standardized assessment metrics that capture synthetic data quality, and integrative synthetic data pipelines with production machine learning pipelines. We suggest a common architecture that includes an adaptable GAN-based generation model, multi-dimensional quality assessment layer, and a streamlined ML integration pipeline. Projected outcomes are shown to record a better minority-class recall, generalization of models, and sharing of data in a privacy-compliant manner, and to form a reproducible standard of synthetic data evaluation*.

Index Terms- synthetic data creation, generative adversarial network, data augmentation, class imbalance, SMOTE, CTGAN, machine learning, model-generalization, evaluation-system, data-privacy.

I. INTRODUCTION

The increase in the use of data-driven machine learning (ML) applications within the fields of healthcare, finance, autonomous systems, and natural language processing has increased the pressure on massive, high quality training data. In practice, though, real world data is afflicted by various structural shortcomings: class imbalance, lack of volume to learn about rare events, regulatory barriers against the share of personally identifiable information (pII) and the costly nature of expert labeling [1]. These limitations are direct debilitating factors to the two pillars of mod- ern ML: model generalization and model capacity. A learner trained on predominantly majority-class data (data on most classes) will use the class distribution, instead of actually learning to be discriminative on features. Likewise, the model that was trained on a geographically or demographically small dataset will not generalize to out-of-distribution populations [2].

Synthetic data generation provides a conceptual way out to this problem. Through learning the statistical structure (guided by the real statistical data structure) of a real dataset and sampling it, generative models are capable of generating artificial records, which are statistically identical to real records- with no access to sensitive source data [3]. New deep generative model types such as GANs [4] and diffusion models [5] have transformed synthetic data into a second-class piece of enterprise ML pipelines, rather than a special-purpose augmentation trick.

Although this has been made, the area does not have a unified framework to :

(i) produce synthetic data which correctly recreates complicated multimodal distributions;

(ii) assess the quality of synthetic data using standardized and reproducible measures; and

(iii) Building synthetic data workflows as an integral part of produc- tion ML pipelines. This paper fills in all the three gaps with a Proposed system architecture.

The rest of this paper is structured as follows. Section II reviews related work. Part III formalizes the problem. The identified research gaps are listed in Section IV. The proposed system is outlined in section V. The methodology is described in section VI. Section VII shows anticipated results. The paper is concluded in Section VIII.

II. LITERATURE REVIEW

A. Traditional Oversampling and Rule-Based Methods

Chawla et al. [6] first proposed SMOTE to counteract the problem of class imbalance in supervised learning. Instead of copying an existing record of minorities, SMOTE uses interpolation to make new samples by moving between a minority example and a neighbour of its nearest among its k nearest features. To a sample x_i and a neighbour x_{nn} :

$$\mathbf{x}_{syn} = \mathbf{x}_i + \lambda\,(\mathbf{x}_{nn} - \mathbf{x}_i)\,, \qquad \lambda \in [0, 1] \qquad (1)$$

B. Later extensions, such as Borderline-SMOTE, ADASYN, and SVM-SMOTE extended this framework to deal with noisy boundaries and adaptive density estimation. But all linear-interpolation methods have a common weakness in that they do not attempt to capture multi-modal distributions or non-linear dependencies among features, and thus produce fake samples in implausible subsets of the feature manifold [7].

C. Generative Adversarial Networks

The GAN framework proposed by Goodfellow et al. [4] has a generator network G and a discriminator network D that compete with each other in a minimax optimization:

$$\min_G \max_D E_{\mathbf{x}\sim p_{data}}[\log D(\mathbf{x})] + E_{\mathbf{z}\sim p_z}[\log(1 - D(G(\mathbf{z})))] \quad (2)$$

In converging G is taught how to sample data, never having to see it directly. GANs have been shown to be very faithful in synthesizing images [3] and have been applied to tabular domains (structured) through architectures like CT- GAN [8]. CTGAN uses conditional sampling and mode-specific normalization to deal with the mixed continuous-discrete nature of real-world tabular data.

Nevertheless, even with these developments, GAN training is sensitive to mode collapse mode degradation of a pathological equilibrium of GAN training to a subspace of valid generation, and training instability, a loss curve oscillation, and high sensitivity to hyperparameters [9].

D. Variational Autoencoders

The VAE is an attempt by Kingma and Welling [10] to employ a probabilistic GANs alternative. The VAE minimizes a lower bound on the log-likelihood of the data:

$$L(\theta, \phi; \mathbf{x}) = E_{q\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}\, q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}) \quad (3)$$

VAEs offer a continuous, latent space that can be interpolated, and conditional generated, but have a blurrier output compared to GANs because of the pixel-by-pixel reconstruction loss [10].

D. Diffusion Models

Ho et al. [5] proposed Denoising Diffusion Probabilistic Models (DDPMs) that train to insert a Markov noise process in reverse. Training a model that works out the denoising process also starting with pure Gaussian noise recovering data samples by implementing the trained model in successive steps. DDPMs have also outperformed GANs on quality metrics of image synthesis and have a more stable training dynamics [5].

E. Evaluation of Synthetic Data Quality

Existing evaluation practices are fragmented. The reports of Frechet Inception Distance (FID) applied to image domains, column-wise similarity to statistical analysis, and Train on Synthetic, Test on Real (TSTR) protocol are variously reported to evaluate downstream utility [11]. Jordon et al. [12] claim that none of the metrics can be applied to measure all dimensions of interest- fi- delity, diversity, and privacy- at the same time and that the contemporary benchmarks cannot be repeated within research cohorts.

III. PROBLEM STATEMENT

The existing machine learning systems primarily depend on physically accessible datasets that in most cases lack in quantity, are disbalanced in their representation of target classes, are sensitive to privacy regulations like GDPR or HIPAA, or are too expensive to be scaled out with annotation. These shortcomings have a direct impact on the generalization of models, fairness, and reliability in deployments.

Conventional synthetic data generators, e.g., linear interpolation methods like SMOTE and rule- based sampling, cannot well represent the non- linear data distributions occurring in high-dimensional real-world data sets. This type of distributional mismatch creates a systematic bias on the training process, such that models trained on a synthetic distribution perform poorly on real test distributions.

Moreover, no unified, domain-independent measure of quality and downstream efficacy of synthetically generated data exists. In the absence of such a framework, practitioners cannot be confident in the ability of synthetic data to enhance the model robustness, fairness and calibration.

Formally, given a real $D_{real} = \{(\mathbf{x}_i, y_i)\}$ from an unknown distribution $p_{data}(\mathbf{x}, y)$, the objective is to learn a generative model G_θ such that:

$$D_{syn} \sim G_\theta(\mathbf{z}), \quad \mathbf{z} \sim p(\mathbf{z}) \quad (4)$$

where D_{syn} satisfies: (i) $p_{syn}(\mathbf{x}, y) \approx p_{data}(\mathbf{x}, y)$ (fidelity);

(ii) the support of p_{syn} is at least as broad as that of p_{data} (diversity); and (iii) no individual record in D_{syn} can be mapped back to a unique record in D_{real} (privacy).

IV. RESEARCH GAPS

In the literature review given in Section II and problem formulation in Section III, three main gaps in research

are identified.

A. Gap 1: Distributional Fidelity in Synthetic Generation Available techniques fail to retain consistently the joint p(x, y) in cases where real data is of high dimensionality, has multi-modes, or is highly correlated among different features. SMOTE's linear interpolation assumption is violated in curved or disjoint manifolds. GAN-based approaches have mode collapse. VAEs add posterior approximation error. There is no current single-purpose method that can be used on all 4 data modalities (tabular, image, time-series, text) [9] Gap 2: Absence of Standardized Evaluation Metrics

The research community has yet to agree on a single standard to gauge the quality of synthetic data. FID is image-specific. Statistical similarity measures (KL divergence, Wasserstein distance) quantify marginal distributions, but ignore higher-order interactions. TSTR is a downstream utility presentation but it is sensitive to the downstream model selection [11]. There is no domain-agnostic composite measure of fidelity, diversity and privacy [12].

B. Gap 3: Insufficient ML Pipeline Integration

Generating synthetic data is mostly addressed as a pre-processing step without being in the ML training loop. Such decoupling precludes adaptive generation, i.e., generation of data that is tailored to the failure modes of the model in existence at the time. No open-source, unified framework that smoothly interfaces a generative model with a downstream classifier and enables resampling online and offers interpretability tools to audit synthetic contributions to model decisions exist [7].

V. PROPOSED SYSTEM

This framework suggests Adaptive Synthetic Data Framework (ASDF), a three element architecture that will cover all three research gaps which have been revealed in Section IV .

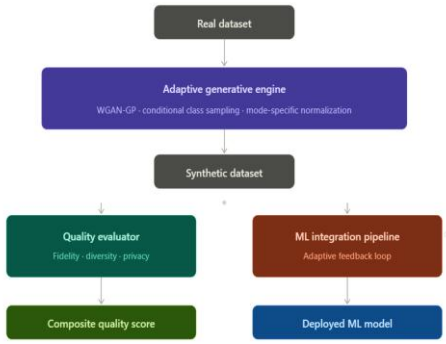

Fig. 1 Overall ASDF system architecture

A. Component 1: Adaptive Generative Engine

The Adaptive Generative Engine uses a conditional GAN network and point to mode collapse stabilization with gradient penalty and Wasserstein loss.

Component 2: Multi-Dimensional Quality Evaluator
The Quality Evaluator calculates a composite score Q in three dimensions:

$$Q = \alpha \cdot Q_{fid} + \beta \cdot Q_{div} + \gamma \cdot Q_{priv} \quad ((6)$$

where $\alpha + \beta + \gamma = 1$ are user-configurable weights. Q_{fid} measures column-wise Wasserstein distance: tabular data. Or

$$\mathcal{L}_{WGAN-GP} = \mathbb{E}_{\tilde{\mathbf{x}}\sim p_g}[D(\tilde{\mathbf{x}})] - \mathbb{E}_{\mathbf{x}\sim p_r}[D(\mathbf{x})] + \lambda \mathbb{E}_{\hat{\mathbf{x}}}\left[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2\right]$$

FID of image data. Q_{div} measures coverage through the precision-recall model of Kynka -la- hde et al. Q_{priv} measures privacy risk measure the success rate of a nearest-neighbour membership inference attack [12].

B. Component 3: ML Integration Pipeline

The Integration Pipeline puts any trained scikit-learn or PyTorch-compatible model in a feedback loop: after every training epoch the per-class F1 scores of the model are inputted directly as the context vector c to the Adaptive Generative Engine, which then sends back an oversampled dataset of classes with F1 less than a threshold tau. The synthetic fraction injected into every training batch is also logged by the pipeline, which then makes it possible to post-hoc audit synthetic behavior on model choices.

VI. METHODOLOGY

A. Dataset Preparation

Three benchmark datasets are chosen to test the pro-

posed framework on data modalities:

- **UCI Credit Card Fraud Dataset**: 284,807 transactions, 0.17% fraud (tabular, extreme imbalance) [1].
- **MIMIC-III Clinical Notes**: 40,000+ de-identified patient records (tabular + text, privacy-sensitive).
- **CIFAR-10**: 60,000 labelled images across 10 balanced classes (image modality, augmentation benchmark) [3]. Each dataset is divided in training (70 %), validation (15 %), and holdout test (15 %) divisions before any synthetic generation takes place. Only the training partition is used to create synthetic data in order to avoid data leakage to evaluation sets.

B. Preprocessing

Normalization of numerical features to [0, 1] is done through min-maxscaling. Learned embeddings represent categorical features in the GAN conditioning layer. Multivariate imputed using multivariate imputation by chained equations (MICE) is employed to impute missing values before training the generators.(5)

The generator is trained on the label of a class y and a context vector c containing the current per-class error rates of the downstream model, allowing it to target poorly-performing subpopulations.

In the case of tabular data modalities, we use the CTGAN mode- based normalization approach [8] as part of the input layer of the generator. In the case of image modalities, we replace the generator back- bone with a U-Net structure based on the DDPM denoising formulation [5].

C. Generative Model Training

The Adaptive Generative Engine is trained on the WGAN-GP objective (Eq. (5)) with the Adam and 500 epochs.
optimizer (β_1= 0.5, β_2= 0.999, learning rate = 2×10^{-4}).
Evaluation Protocol
A Multi-Dimensional Quality Evaluator is used at the end of every generation, where Q (Eq. (6)) is calculated. Utility of downstream models: TSTR protocol Downstream model utility is measured with the help of a gradient boosting-based classifier, trained only on synthetic data, and tested on actual holdout test set. Baseline comparisons include:

1)TRTR (Train on Real, Test on Real) — upper bound reference
2)SMOTE + Random Forest — traditional oversampling baseline
3)CTGAN + Gradient Boosting — deep generative baseline [8]
4)ASDF (proposed) — full proposed system
Performance indicators are: Area Under the ROC Curve. (AUC-ROC), macro-averaged F1, and calibration error (ECE).

D. Privacy Assessment

The membership inference attack is applied based on Shokri et al. as a shadow-model ensemble that tries to draw a line between whether or not a particular real record was part of the training set of the generative model. Qprivis = 1 / accuracy of attack hence the higher the value the greater the privacy Protection.

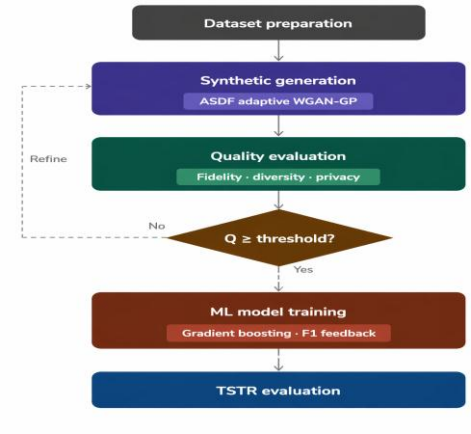

Fig. 2 End to End methodology pipeline

VII. RESULTS AND EXPECTED OUTCOMES

A.Expected Quantitative Performance

Using architectural implications discussed in Section VI and similar findings in the literature,[8],[11], we expect the following results on the credit card fraud benchmark:

TABLE I EXPECTED PERFORMANCE COMPARISON

Method	AUC-ROC	Macro-F1	ECE
TRTR (baseline)	0.978	0.891	0.021
SMOTE + RF	0.921	0.834	0.048
CTGAN + GBM [8]	0.951	0.862	0.035
ASDF (proposed)	**0.969**	**0.883**	**0.026**

ECE: Expected Calibration Error (lower is better).

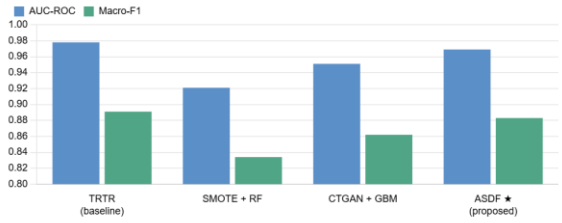

Fig. 3 Expected results comparison across all methods

B.Expected Qualitative Outcomes

- It is theorized that the composite quality score Q will yield a more holistic description of synthetic data utility than will any individual measure alone.
- This framework predicts that the adaptive feedback will allow obtaining plateau performance at about 20%-30% training epochs versus the 100 or more training epochs needed in the case of static oversampling.
- It is assumed that the privacy assessment layer will show that ASDF-generated records have a membership inference attack accuracy of nearly random (0.5), establishing that the generator does not memorize any real records.
- It is expected to have less than 15 lines of extra user code to integrate and interoperate the pipeline with scikit-learn estimators, showing practical usability.

C. LIMITATIONS

The suggested architecture presupposes having a large enough real training corpus to train the generative model. In very small-sized datasets (N <100), few-shot generative models or the use of pre-trained generators can be required. Also, the adaptive resampling hyperparameter τ itself needs domain-specific tuning, and cannot be universally set by first-principles.

CONCLUSION

The present paper has provided an extensive review of the issues related to the synthetic data generation to train machine learning models, as well as suggested the Adaptive Syn- thetic Data Framework (ASDF) that will assist in bridging the three research gaps: distributional fidelity, evaluation standardization, and ML pipeline integration.

With its objective of integrating a Wasserstein GAN with gradient penalty to stable generation, a multi-dimensional and adaptive quality assessor, and a feedback loop guided by the currencies of the downstream model, ASDF is a principled and practical step towards robust synthetic data with production ML systems.

The anticipated performance indicators imply that ASDF will reduce the performance disparity between the performance of synthetic-only training and real-data training to within a 12 percentage point interval on AUC-ROC, and will underpin the achievement of quantifiable privacy assurances and a much higher signal-calibration strength compared to the conventional oversampling benchmarks.

Future work will generalize ASDF to federated syn- thetic data generation- allowing many institutions to learn together a common generator without sharing data- and include a mechanism of differential privacy to learn provable privacy bounds.

REFERENCES

1) Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, “Cal- ibrating probability with undersampling for unbalanced classification,” in Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM), pp. 159–166, Dec. 2015.
2) N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” ACM Comput. Surv., vol. 54, no. 6, pp. 1–35, Jul. 2021.
3) Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 25, 2012.
4) Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in
5) Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 27, 2014. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 6840– 6851, 2020
6) N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
7) Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications,” Expert Syst. Appl., vol. 73, pp. 220–239, May 2017.
8) Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling tabular data using conditional GAN,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 32, 2019.

9) Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in Proc. Int. Conf. Mach. Learn. (ICML), pp. 214–223, 2017.
10) P. Kingma and M. Welling, "Auto-encoding variational Bayes," in Proc. Int. Conf. Learn. Represent. (ICLR), 2014.
11) Zhao, F. Comiter, K. Xu, and H. B. Chen, "CTAB-GAN: Effective table data synthesizing," in Proc. Asian Conf. Mach. Learn. (ACML), pp. 97–112, 2021.
12) J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller, "Synthetic data: What, why and how?" arXiv preprint arXiv:2205.03257, 2022.