IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Synthetic Data Generation For Rare Event Detection Using Class-Conditional Diffusion Models

Heena Kouser
Department of CSE
KVGCE, Sullia,
Dakshina Kannada, INDIA

Bhavya P S
Department of CSE
KVGCE, Sullia,
Dakshina Kannada, INDIA

Balapradeep K N
Department of CSE
KVGCE, Sullia,
Dakshina Kannada, INDIA

Sindhu Venkatesh
Department of CSE
KVGCE, Sullia,
Dakshina Kannada, INDIA

Abstract

Across several recent studies, diffusion models [10] have become leading approaches for generating high-quality synthetic data, especially for complex, limited, or sensitive datasets. These works emphasize the increasing demand for effective data generation and imputation methods in areas such as healthcare, finance, and rare-event detection, where real-world data is often scarce, imbalanced, or constrained by privacy concerns. Diffusion-based frameworks [11][15]—including TabDDPM [17], transformer-based architectures, and privacy-preserving federated diffusion systems [25]—have shown significant advantages over conventional GAN and VAE models [2], providing more stable training, higher-quality sample generation, and better support for mixed-type tabular data.

These models are capable of handling tasks such as Rare– event augmentation, missing data imputation. and privacy preservation. Compliant financial data synthesis, and pandemic related medical image generation [8] Surveys and systematic reviews further emphasize diffusion models' growing role in addressing challenges like class imbalance, multimodal feature distributions, missing values, and privacy risks. Overall, the combined body of work shows that diffusion models are reshaping synthetic data generation across modalities—enabling more accurate, fair, and privacy-safe machine learning in situations where data is scarce, sensitive, or difficult to obtain.

Keywords— Diffusion models, synthetic data generation, rare event detection, imbalanced datasets, tabular data, data augmentation, differential privacy, federated learning, conditional generation, data imputation, transformer models, privacy- preserving machine learning, financial data, healthcare data, anomaly detection.

financial defaults are of disproportionate importance despite representing a tiny fraction of real-world.

Introduction

information. The minority (rare) class is frequently less than 1% of all observations in highly imbalanced tabular and EHR datasets,

which causes traditional machine learning (ML) models to concentrate on majority
patterns and perform badly on such uncommon but significant occurrences. Access to rich rare-event data is further restricted, particularly in healthcare and finance, by privacy restrictions (e.g., GDPR, HIPAA) [16], restricted data sharing across institutions, and costly annotation costs.

A principled solution to these limitations is provided by synthetic data generation: a generative model learns an approximation of the true data distribution and then creates artificial records that resemble the statistical and structural characteristics of real data but are more freely shared and altered. Class imbalance is only slightly alleviated by early synthetic tabular approaches and conventional oversampling techniques, which also struggle with complex connections and diverse characteristics. Deep generative models, especially diffusion models and their class conditional variations, have been shown in recent research to produce high-fidelity tabular and EHR data, enhance rare-event detection performance, and facilitate privacy-aware data sharing.

Background

I. Rare Events and Class Imbalance

With an emphasis on tabular and EHR domains, this study examines **"Synthetic Data Generation for Rare Event Detection using Class-Conditional Diffusion Models."**[15] The objectives are to give a thorough but understandable review of the problem setting, highlight important generating approaches with a focus on diffusion models, talk about evaluation and difficulties, and suggest possible avenues for further study.

Uncommon Occurrences and Class Disparities Think about supervised learning using labels $y \in \{0,1\}$ and features $\mathbf{x} \in \Psi d$ (or mixed-type vectors), where $y = 1$ indicates an uncommon occurrence (e.g., sepsis, fraud, machine failure). In large-scale EHR or financial datasets, the class prior $\pi = P(y = 1)$ is usually very small, frequently less than 0.01. The formula for the empirical risk of a classifier $f\theta$

and high dimensional EHRs or financial tables, and they usually assume simple distributions or demand strong parametric assumptions. III. Deep Generative Models for Variational Autoencoders (VAEs) of Tabular

under a loss Ψ is $\mathscr{R}(f\theta) = (1 - \pi)\ \mathbb{E}.\ [\ (f\theta(\mathbf{x}),0) \neq y = 0] + \pi\mathbb{E}$When π is small, the rare class makes a negligible contribution to the optimization goal $[\ (f\theta(\mathbf{x}),1) \neq y = 1]$. Because of this, unaltered training frequently results in low recall and poor calibration for the minority class, favoring overall accuracy and majority-class performance.

$\mathcal{R}(f_\theta) = (1 - \pi)\ \mathbb{E}[\ell(f_\theta(\mathbf{x}),0) \mid y = 0] + \pi\ \mathbb{E}[\ell(f_\theta(\mathbf{x}),1) \mid y = 1],$

Class weighting, cost-sensitive losses, threshold tuning, random over/under-sampling, and artificial oversampling techniques like SMOTE, Border line SMOTE, and ADASYN are examples of conventional techniques to reduce imbalance. These methods can enhance performance, but they typically do not model the true conditional distribution $p\ (\mathbf{x} \neq y = 1)$. They may also fail when rare patterns are highly structured, nonlinear, or high-dimensional, as is frequently the case with tabular and EHR data that combine continuous labs, demographics, codes, and time-related features.

II. Synthetic Data for Tabular and HER Data

The goal of synthetic data generation is to train a generative model $p\theta(\mathbf{x})$(or $p\theta(\mathbf{x} \neq y)$) so that samples $\tilde{\mathbf{x}} \sim p\theta$ preserve pertinent features of the actual data distribution $p$data($\mathbf{x}$). Three primary objectives in tabular and EHR situations are:

• Data augmentation: To enhance model training and evaluation, produce more samples, such as rare-event data.

• Privacy-preserving sharing: Provide artificial datasets that mimic actual data while lowering the risk of re-identification, allowing for more widespread data access and method advancement.

• Imputation and scenario analysis: Create data under predetermined conditions (e.g., certain covariates or treatments) to fill in missing values or model "what-if" scenarios. Gaussian mixture models, copulas, Bayesian networks, and traditional oversampling strategies are examples of conventional probabilistic models used for tabular synthesis. However, these approaches frequently struggle with very heterogeneous

and EHR Data With a variational encoder $q\phi(\mathbf{z} \neq \mathbf{x})$, VAEs introduce continuous latent variables $\mathbf{z}$ and model $p\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p\theta(\mathbf{x} \neq \mathbf{z})$. The evidence lower bound (ELBO) is maximized through training:

$\mathcal{L}$VAE = $\mathbb{E}q\phi(\mathbf{z} \bullet \mathbf{x})[\log p\theta(\mathbf{x} \neq \mathbf{z})] - \mathrm{KL}(q\phi(\mathbf{z} \neq \mathbf{x}) \times p(\mathbf{z}))$. Using specific likelihoods and encoders, tabular VAEs like TVAE and GOGGLE extend this paradigm to mixed-type information (continuous, categorical).

$\min \max \mathbb{E}_{\mathbf{x}\sim p\text{data}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$

$G \quad D$

$\min \max \mathbb{E}\mathbf{x}\sim p\text{data}[\log D(\mathbf{x})] + \mathbb{E}\mathbf{z}\sim p(\mathbf{z})[\log(1 - D(G(\mathbf{z})))]$ G D Tabular-specific GANs such as CTGAN and CTABGAN introduce conditional generation, modespecific normalization, and tailored handling of categorical variables. In EHR settings, models like medGAN and TableGAN generate sequences of medical codes and continuous attributes to support synthetic record sharing and augmentation. While GANs can produce high-fidelity samples, they are strong parametric assumptions, and they often struggle with highly heterogeneous and high dimensional EHRs or financial tables.

III. Deep Generative Models for Tabular and EHR Data

Autoencoders that vary (VAEs) With a variational encoder $q\phi(\mathbf{z} \neq \mathbf{x})$, VAEs introduce continuous latent variables $\mathbf{z}$ and model $p\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p\theta(\mathbf{x} \neq \mathbf{z})$. The evidence lower bound (ELBO) is maximized through training:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z})).$$

Using specific likelihoods and encoders, tabular VAEs like TVAE and GOGGLE extend this paradigm to mixed-type information (continuous, categorical). Although they can produce realistic synthetic tabular data and frequently beat traditional methods in downstream machine learning tasks, they have a tendency to produce distributions that are too smooth and may underrepresent tail and sharply multimodal behaviors, which are important for rare events.

IV. Generative Adversarial Networks (GANs)

In a minimax game, GANs pit a discriminator D against a generator G:

$\min \max \mathbb{E}\mathbf{x}\sim p\text{data}[\log D(\mathbf{x})] + \mathbb{E}\mathbf{z}\sim p(\mathbf{z})[\log(1 - D(G(\mathbf{z})))]$

Conditional generation, mode-specific normalization, and customized treatment of categorical variables are introduced by tabular-specific GANs like CTGAN and CTABGAN. To facilitate the exchange and enhancement of synthetic records in EHR contexts, methods such as medGAN and TableGAN provide sequences of medical codes and continuous data. Although GANs are capable of producing high-fidelity samples, they are known to experience mode collapse and unstable training. Additionally, they have the potential to memorize training records, which could pose a privacy concern in sensitive sectors.

V. Diffusion Models and Score-Based Models

A typical diffusion model q, $\beta t\mathbf{I})$, t = 1,…, T, with a variance schedule $\{\beta t\}Tt$=1. The marginal distribution that results is: q, t, where $\alpha^- t = \prod s$=1(1 − βs).

$(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t\mathbf{I}), t = 1, \ldots, T,$

with a variance schedule $\{\beta_t\}^T{}_{t=1}$. The resulting marginal distribution is:

$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I})$,

where $\alpha^-{}_t = \prod_{s=1}^{t}(1 - \beta_s)$.

The model for the opposite process is:

$p\theta(\mathbf{x}t-1 \neq \mathbf{x}t) = \mathcal{N}(\mathbf{x}t-1; \boldsymbol{\mu}\theta(\mathbf{x}t, t), \sigma t2\mathbf{I})$, and the network is trained to anticipate the noise that was

added at each stage. A typical example of a simple loss is:

Managing Different Feature Types

Continuous variables (such as lab values and sensor readings), categorical variables (such as diagnosis codes and demographic categories), ordinal scores, and occasionally temporal indices are all present in realistic tabular and EHR data sets. For such data, diffusion models use either:

- Dual diffusion: A joint denoiser network that predicts both continuous noise and categorical logits, with distinct Gaussian diffusion chains for continuous features and multinomial (categorical) diffusion for one-hot encoded categorical data.
- Latent diffusion: Mixed-type features are first mapped to a continuous latent space by a VAE or comparable encoder, where they are learned; latent samples are then decoded back to discrete and continuous features.

A typical dual-diffusion model is TabDDPM, which employs a multi-layer perceptron denoiser after applying multinomial diffusion to categorical columns and Gaussian diffusion to normalized continuous columns. Both unconditional and class-conditional generation are supported by TabDDPM; for classification datasets, it uses embedding vectors to condition on labels, allowing for targeted minority oversampling.

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\mathbf{x}_0, t, \boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\left[\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, t\right) \|_2^2 \right].$$

The ability to produce synthetic data conditioned on the rare class label—that is, from $p\theta(\mathbf{x} \neq y = 1)$—is crucial for rare-event detection. Label information is incorporated into the denoising network by class-conditional diffusion models through: • Including label embeddings in the intermediate layers or network input. Using classifier guidance or classifier-free guiding, This viewpoint is extended to continuous-time SDEs by score-based models, which solve a reverse-time SDE or ODE to sample new points after learning the score $\nabla\mathbf{x}\log pt(\mathbf{x})$ at various noise levels. These models have demonstrated state-of-the-art performance in several generative tasks, such as images and increasingly tabular data, and are especially expressive.

VI. Diffusion Models for Tabular and EHR Data

Mixed-type tabular datasets typically contain continuous numerical values (e.g., sensor readings), categorical variables (e.g., diagnosis codes or demographic groups), ordinal scores, and sometimes temporal indices. To model such heterogeneous data, diffusion-based approaches generally follow one of two strategies:

- **Dual diffusion:** Uses separate diffusion processes for different data types — Gaussian diffusion for continuous variables and multinomial diffusion for one-hot encoded categorical variables. A shared denoising network jointly predicts continuous noise and categorical probabilities (logits).
- **Latent diffusion:** First applies a Variational Autoencoder (VAE) or a similar encoder to transform mixed-type data into a continuous latent representation. Diffusion is then performed in this latent space, and the generated latent samples are decoded back into both continuous and discrete features.

VII. Class-Conditional Diffusion for Rare Events

which combines conditional and unconditional predictions to direct the reverse process toward a desired label.

Classifier-free guidance involves training a single network to produce both conditional and unconditional noise predictions, then interpolating between them using a guiding weight w.

$$\epsilon\text{guided}\theta(\mathbf{x}t, y, t) = \epsilon\text{uncond}\theta(\mathbf{x}t,\ t) + w(\epsilon\theta\text{cond}(\mathbf{x}t, y, t) - \epsilon\text{uncond}\theta(\mathbf{x}t, t)).$$

The model can be utilized as a potent over sampler for training rare-event detectors by choosing the rare class label $y = 1$ and setting $w > 1$, which emphasizes the creation of synthetics congruent with the rare-event manifold.

VIII. Related Works Surveys and Systematic Reviews

The progress of diffusion-based generative approaches and its applications in structured data synthesis, imputation, augmentation, and privacy-preserving data generation are reviewed in Li et al.'s thorough analysis of diffusion models for tabular data.

IX. GAN/ VAE-Based Synthetic Tabular and EHR Frameworks

CTGAN and CTAB-GAN are popular tabular GAN frameworks that leverage conditional generation and mode-specific normalization to handle mixed-type information. They have been used for augmentation and privacy-preserving release in financial, marketing, and healthcare-style tables. Although VAEs like TVAE and GOGGLE offer probabilistic latent-feature models, can be simpler to train, and offer some quantification of uncertainty, they frequently produce "blurred" or averaged uncommon patterns that are less appropriate for modeling extreme behavior. From copulas, Bayesian networks, and SMOTE to VAEs, GANs, and diffusion-based techniques, medGAN and TableGAN are utilized in healthcare to generate discrete EHR records. Data augmentation and oversampling, data imputation, trustworthy synthesis (privacy, fairness), and anomaly detection are among the applications for which they classify diffusion models. They also highlight important issues such mixed feature types, missingness, tiny datasets, and domain-specific limitations.

MedGAN and TableGAN are used in healthcare to create discrete EHR records using copulas, Bayesian networks, SMOTE, VAEs, GANs, and diffusion-based methods. They categorize diffusion models for applications such as data augmentation and oversampling, data imputation, trustworthy synthesis (privacy, fairness), and anomaly detection. Important problems like heterogeneous feature types, missingness, small datasets, and domain-specific constraints are also highlighted. show that synthetic EHR sharing is feasible, but further research has revealed problems with mode coverage and privacy leakage.

Diffusion models may be more resilient in the face of data scarcity and highly unbalanced distributions, particularly for structured data, according to comparative research on small-sample machine learning.

X. Diffusion Models for Tabular and EHR Data

TabDDPM, CoDi, TabSyn, TabDiff, and Forest. A family of diffusion models tailored to tabular data is represented by Diffusion, CDTD, and TabUnite. They vary in their application focus (augmentation vs. imputation vs. privacy), denoiser architecture (MLP vs. transformer vs. GNN-based), and how they represent mixed types (dual diffusion vs. latent diffusion). Diffusion-based models frequently beat CTGAN and TVAE in terms of downstream classifier performance (ML efficiency) [17] and distributional fidelity (Wasserstein, Jensen–Shannon, correlation measures) across common

benchmarks, including Adult, HELOC, Otto, Cardio, and others.

By modeling the conditional distribution of masked features given observed features and labels and consistently outperforming GAN, VAE, and previous diffusion baselines in ML efficiency, statistical similarity, and membership inference attack resistance, MTabGen shows that a transformer-based diffusion model with dynamic masking can concurrently solve data imputation and synthetic generation.

XI. Privacy-Preserving and Federated Diffusion

Noisy gradient updates are combined with federated training across institutions in differentially private and federated diffusion frameworks, such as DP-Fed-FinDiff [19] and related cross-silo tabular diffusion models. These studies demonstrate that diffusion models can offer formal or empirical privacy guarantees while maintaining sufficient structure for subsequent tasks and can be trained in a distributed manner without centralizing raw data. They also draw attention to the difficulty of striking a balance between the requirement for high fidelity minority-class modeling and stringent DP constraints, since the additional noise frequently has a disproportionate effect on rare-event accuracy.

Review Methodology

This review focuses on techniques that:

- Produce tabular or EHR data that is synchronized.
- Use conditioning, oversampling, or anomaly modeling to deal with uncommon occurrences or extreme class imbalance.
- Make use of contemporary deep generative models, emphasizing class conditional variations and diffusion.

Papers are grouped according to:

- Application role: o Rare-event oversampling and data augmentation.
- Unified generation imputation and imputation.
- Reliable synthesis (fairness, privacy, federated learning).
- The use of generative scores or likelihoods for anomaly and rare-event detection.
- Data domain: o General tabular data (e.g., sensors, credit risk, finance).
- Healthcare/EHR data (risk prediction, multimorbidity, etc.).

The study takes into account the data representation (feature types and preprocessing) for each approach. model architecture (VAE, GAN, diffusion, hybrid), evaluation protocol (utility, fidelity, privacy, and, if relevant, fairness), and conditioning techniques (class labels, side information, federated setup, DP noise). The focus is on how these approaches compare in real-world scenarios and how they can facilitate rare-event identification through the creation of synthetic data.

Key Challenges

Heterogeneous And Complex Feature Distributions

Due to the intrinsic heterogeneity of tabular and EHR data, modeling is further complicated by ordinal and temporal properties; continuous features might be skewed, heavy-tailed, and multimodal; and categorical features can have extremely imbalanced

In comparison to GAN- and VAE-based synthetic EHR generators, Kita et al.'s DDPM-based architecture for structured EHR tables shows greater integrity, better preservation of pairwise feature correlations, and lower membership-inference risk. Additionally, they demonstrate the potential of diffusion-based EHR synthesis for exploratory analysis and unsupervised learning by demonstrating how clustering synthetic EHRs using Dirichlet process mixture models recovers clinically significant multimorbidity patterns.

category frequencies. Diffusion models have to either acquire representations that align these many modalities in a common continuous space or create distinct noise processes for each form of data. When

Extreme Rarity and Tail Modelling

The knowledge regarding unusual events is restricted to a few examples, even in the case of class-conditional diffusion. Particularly when privacy constraints are also applied, the model may overfit to certain rare-event cases, underfit tails, or oversmooth sharp borders. It is necessary to carefully balance model capacity, regularization, training objectives, and, in certain situations, data augmentation or domain knowledge to enrich the rare-event manifold in order to capture the shape of $p(\mathbf{x} \neq y = 1)$. Assessment: Privacy, Fidelity, and Utility.

Three primary dimensions are involved in the evaluation of synthetic data:

- Utility (ML efficiency): Measure measures such as AUC-PR, F 1 for the rare class, and calibration scores; train classifiers or regressors on synthetic or augmented data and test on real hold-out data.
- Fidelity (distributional similarity): Use correlation matrices for continuous–continuous, categorical–categorical, and mixed pairs as well as distributional distances (Wasserstein, Jensen–Shannon, Kolmogorov–Smirnov) to compare actual and synthetic data.
- Privacy: Use membership-inference attacks, attribute-disclosure attacks, and, when appropriate, differential privacy constraints on training to assess risk.
- Trade-offs between utility, fidelity, and privacy—particularly for rare events—can

Discussion and Future Trends

Diffusion models, particularly class-conditional and tabular/EHR-specific variations, appear to provide substantial benefits for synthetic data generation in rareevent detection tasks, according to the reviewed literature. They offer versatile conditioning

synthetic data is handled incorrectly, it may appear somewhat realistic yet miss rare configurations and cross-feature dependencies that are essential for rare-event identification.

be obscured by the use of disparate metrics, datasets, and baselines across articles, making direct comparison difficult.

Trustworthiness, Fairness, and Deployment

Synthetic rare-event data deployment in the real world, particularly in healthcare and finance, necessitates confidence in more than just predicted performance. Models need to guarantee:

- Privacy compliance: No excessive disclosure of personal data, especially for vulnerable and minority populations.
- Fairness and bias control: Biases pertaining to socioeconomic or demographic characteristics shouldn't be amplified or introduced by synthetic generation.
- Domain validity: Synthetic patterns must respect clinical or business limitations and make sense to domain experts.

Although diffusion models can incorporate DP and federated learning and seem less vulnerable to direct memorization than GANs, formal examination of fairness and domain validity in rare-event circumstances is still scarce.

Computational Cost and Practicality

Diffusion models are more computationally expensive than some GAN/VAE techniques since they often need numerous denoising steps for every generated sample.

Efficient samplers, distillation, or approximation techniques are required for applications that need large-scale synthetic datasets or near-real-time creation (e.g., online augmentation in streaming systems).

techniques that directly support targeted rare-event oversampling, improved coverage of complicated distributions, and more stable training than GANs. The capacity to represent conditional distributions and fine-grained feature dependencies is further improved by transformer-based denoisers and conditioning attention in models such as MTabGen. Work on federated and DP-enabled diffusion, on the other hand, shows a shift toward reliable synthetic data pipelines, in which institutions communicate only synthetic outputs or model parameters instead of raw data, and in which DP noise is employed to reduce memory risk. It is still difficult to precisely adjust these systems to preserve privacy without compromising rareevent fidelity.

In the future, a number of study avenues stand out:

• **Tail- and rare-event-aware diffusion objectives:** Class-balanced losses, tail-focused sampling, or extreme value-inspired regularization could enhance rare-event fidelity and robustness by precisely capturing the tails of $p(\mathbf{x} \neq y = 1)$. Training strategies that explicitly focus on learning and sampling challenges remain in modeling heterogeneous feature spaces.

• **Causally-informed synthetic EHR data:** Decision support requires that synthetic trajectories and outcomes be coherent under interventions (such as treatment modifications), not merely under observational correlations. This could be achieved by integrating causal models with diffusion.

• **Standardized standards:** It would be simpler to evaluate approaches and comprehend trade-offs between utility, fidelity, privacy, and fairness if there were community benchmarks for synthetic rare-event tabular/EHR data with established metrics and datasets.

• **Effective and deployable diffusion models:** Methods like model distillation, sampler acceleration, and architectural optimization can assist bring diffusion-based synthetic data creation into settings with limited time and resources.

Conclusion

In tabular and EHR-based machine learning, synthetic data generation is emerging as a key method for addressing rare-event scarcity, imbalance, and privacy limitations. In terms of fidelity, variety, and downstream event detection performance, diffusion models—especially class-conditional architectures tailored to mixed-type structured data—have become potent generative models that frequently surpass GANs and VAEs. However, developing strong evaluation processes, maintaining fairness, and verifying domain validity are crucial under data and privacy restrictions. In high-stakes industries including healthcare, banking, cybersecurity, and industrial systems, securely and successfully using class-conditional diffusion models for synthetic rare-event data production will depend on resolving these issues.

References

1. Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. -2025 Mahmoud Ibrahima, b,c Yasmina Al Khalild Marcel Breeuwerd , Sina Amirrajabd , Chang Suna,b, , Josien Pluimd, Bart Elenc, Gökhan Ertaylanc, Michel Dumontiera

2. Examine the Role of Generative Adversarial Networks (GANs) and Generative AI for Synthetic Data Generation and Augmentation in Machine Learning. –Volume 13, Issue 12, December 2024

3. Systematic Evaluation of Synthetic Data Augmentation for Multi-class NetFlow Traffic ⋆ Maximilian Wolf1 Dieter Landes1 Andreas Hotho2 Daniel Schlör2 arXiv:2408.16034v1 [cs.CR] 28 Aug 2024 1 Centre for Responsible Artificial Intelligence, University of Applied Sciences and

Arts Coburg, Friedrich-Streib-Str. 2 Coburg, Germany {maximilian.wolf, dieter.landes}@hs-coburg.de 2 Centre for Artificial Intelligence and Data Science, University of Würzburg, Campus Hub land Nord, Emil-Fischer-Straße 50 Würzburg, Germany{hotho,schloer}@informatik.uniwuerzburg.de

4. Efficient Diffusion Models in Medical Imaging: A Comprehensive Review ABDULLAH, James Cook University, Australia TAO HUANG, James Cook University, Australia ICKJAI LEE, James Cook University, Australia EUIJOON AHN∗, James Cook University, Australia arXiv:2505.07866v1 [eess.IV] 9 May 2025

5. A Survey of Synthetic Data Generation for Rare Events JINGYI GU∗, XUAN ZHANG∗, and GUILING WANG, New Jersey Institute of Technology, Newark, NJ, USA arXiv:2506.06380v1 [cs.LG] 4 Jun 2025

6. Generative Artificial Intelligence in Medical Imaging: Foundations, Progress, and Clinical Translation Xuanru Zhou1,2, Cheng Li1, Shuqiang Wang3, Ye Li1, Tao Tan4, Hairong Zheng1, and Shanshan Wang*1 1Paul C. Lauterbur Research Centre for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. 2University of Chinese Academy of Sciences, Beijing, China. 3Research Centre for Biomedical Information Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China 4Faculty of Applied Sciences, Macao Polytechnic University, Macao, China. * Correspondence: Shanshan Wang (ss.wang@siat.ac.cn etic)

7. Conditional SynthData Generation for Robust Machine Learning Applications with Limited Pandemic Data-2023 Hari Prasanna Das1, Ryan Tran1, Japjot Singh1, Xiangyu Yue1, Geoffrey Tison2, Alberto SangiovanniVincentelli1, Costas J. Spanos1 1 Department of Electrical Engineering and Computer Sciences, University of California, Berkeley 2 Division of Cardiology, University of California, San Francisco (UCSF) {hpdas, bobotran, calzoom, xyyue, alberto, spanos}@berkeley.edu, geoff.tison@ucsf.edu

8. A Systematic Review of Rare Events Detection Across Modalities Using Machine Learning and Deep Learning YAHAYA IDRIS ABUBAKAR AND AZNUL QALID MD SABRI 1, ALICE OTHMANI 1, PATRICK SIARRY 2 1, 1Laboratoire Images, Signaux et Systémes Intelligents (LiSSi)–EA 3956, Université ParisEst Créteil (UPEC), 94010 Créteil Cedex, France 2Faculty of Computer Science & Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia Corresponding author: Alice Othmani (alice.othmani@upec.fr)- 2024

9. Differentially Private Federated Learning of Diffusion Models for Synthetic Tabular Data Generation- 2024

10. Diffusion Models for Tabular Data: Challenges, Current Progress, and Future Directions- 2024

11. Diffusion Models for Tabular Data Imputation and Synthetic Data Generation MARIOVILLAIZÁN-VALLELADO, Universidad de Valladolid, Spain and Telefonica Research and Development, Spain MATTE OSAL VATORI, Telefonica Research and Development, Spain CARLOS SEGURA, Telefonica Research and Development, Spain IOANNIS ARAPAKIS, Telefonica Research and Development, Spain arXiv:2407.02549v1 [cs.LG] 2 Jul 2024

12. IJIRSET©2025 | An ISO 9001:2008 Certified Journal | 5437 Synthetic Data Pipelines for Training AI Models in Data-Scarce Domains Danish Reddy Agarampalli

13.Enhancing Small-Data Machine Learning with Generative AI: A Comparative Study of GANs, Diffusion Models, and VAEs Anthony Lawrence Paul- 2022

14. SYNTHETICIMU-BASEDTIME-SERIES DATAGENERATIONFORP2WCRASH DETECTION-2025

15. Innovative synthetic EHR data generation: diffusion models for enhanced privacy and clinical utility in multimorbidity clustering Francis John Kita, Gadde Srinivasa Rao & Peter Josephat Kirigiti -2025

16. Synthetic Data Generation for Training Deep Models to Detect Rare Sensor Failures Author: Akintan Favour, Muhammad Swaileh Alzaidi, Vijetha Ringu, Majdy Eltahir, Florian Jungmann, Tobias Jorg Date: 26th July 2025

17. TabDDPM: Modelling Tabular Data with Diffusion Models Akim Kotelnikov Dmitry Baranchuk Ivan Rubachev Artem Babenko -2022

18. Leveraging synthetic data to tackle machine learning challenges in supply chains: challenges, methods, applications, and research opportunities Yunbo Long, Sebastian Kroeger, Michael F. Zaeh & Alexandra Brintrup -2025

19. Text-Conditioned Diffusion-Based Synthetic Data Generation for Turbine Engine Sensor Analysis and RUL Estimation Luis Pablo Morade-León †, David Solís-Martín, Juan Galán-Páez andJoaquín Borrego-Díaz -2025

20. Synthetic data generation by diffusion models Jun Zhu -2024 Volume 14, Issue 4, April 2025 |DOI: 10.15680/IJIRSET.2025.1404005|

21. Generative models improve fairness of medical classifiers under distribution shifts -2023

22. Enhancing Wearable Fall Detection System via Synthetic Data Minakshi Debnath, Sana Alamgeer, Md Shahriar Kabir and AnneH. Ngu* Department of Computer Science, Texas
State University, San Marcos, TX 78666-4684, USA; stg60@txstate.edu (M.D.); sana.alamgeer@txstate.edu (S.A.); cpi12@txstate.edu (M.S.K.) * Correspondence: angu@txstate.edu

23. Synthetic Tabular Data Generation for Imbalanced Classification: The Surprising Effectiveness of an Overlap Class Annie D'souza*, Swetha M*, Sunita Sarawagi Indian Institute of Technology, Bombay -2024

24. TABDIFF: A MULTI-MODAL DIFFUSION MODEL FOR TABULAR DATA GENERATION -2024

25. Time Auto Diff: Combining Autoencoder and Diffusion model for time series tabular data synthesizing -2024