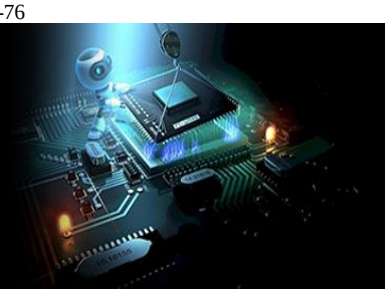
International Journal of Engineering in Computer Science



E-ISSN: 2663-3590
P-ISSN: 2663-3582
www.computersciencejournals.com/ijecs
IJECS 2024; 6(1): 71-76
Received: 10-02-2024
Accepted: 15-03-2024

Srikanth Peddisetti
Senior People Systems Consultant, Parsons Services Company, 100 W Walnut Street, Pasadena, California 91124, USA

Corresponding Author:
Srikanth Peddisetti
Senior People Systems Consultant, Parsons Services Company, 100 W Walnut Street, Pasadena, California 91124, USA

Generative data engineering for the unseen: Harnessing GANs for rare event synthesis

Srikanth Peddisetti

DOI: https://www.doi.org/10.33545/26633582.2024.v6.i1a.177

Abstract

The performance of machine learning systems is highly dependent on the availability of high-quality, balanced datasets. In many real-world domains such as fraud detection, medical diagnosis, and fault prediction, rare events are significantly underrepresented, leading to skewed data distributions and poor model generalization. This paper proposes a novel Generative Data Engineering framework leveraging Generative Adversarial Networks (GANs) to synthesize realistic, high-fidelity data for rare classes. The framework integrates data preprocessing, conditional GAN architectures, and a feedback loop to generate semantically valid and structurally consistent synthetic samples for effective minority class augmentation.

The framework was evaluated on two publicly available datasets: the healthcare domain, using rare cardiac arrhythmia records from the MIT-BIH Arrhythmia Database, and the cybersecurity domain, with minority-class attack instances from the NSL-KDD intrusion detection dataset. Key evaluation metrics included accuracy, minority class recall, F1-score, and detection rate. Results showed a significant increase in recall from 62% to 89% in healthcare and from 68% to 92% in cybersecurity after augmenting with GAN-generated data. Additionally, Fréchet Inception Distance (FID) scores below 20 across both datasets indicate high similarity between real and synthetic data. Ablation studies further demonstrated the critical role of the feedback mechanism, with recall dropping by 15% and 11% when omitted in healthcare and cybersecurity datasets, respectively.

These findings confirm that the proposed Generative Data Engineering framework enhances model sensitivity, generalization, and interpretability while maintaining data realism, offering a scalable solution for rare event modeling in safety-critical domains.

Keywords: Synthetic data, rare events, generative adversarial networks, data imbalance, generative data engineering, anomaly augmentation, conditional GANs

1. Introduction

Data imbalance is a prevalent challenge in machine learning, especially in domains where critical events occur infrequently. These include rare disease diagnosis, financial fraud detection, and network intrusion identification. Standard supervised learning algorithms tend to be biased toward majority classes, leading to poor detection and generalization for rare classes. Traditional techniques like oversampling, SMOTE, and class reweighting often fall short in capturing the true distribution of rare events. These methods frequently introduce artificial patterns that do not exist in the original data or fail to preserve the intrinsic properties of minority class samples. As a result, models trained on such datasets often exhibit poor sensitivity and higher false negatives, which are especially undesirable in critical applications like healthcare and cybersecurity.

Recently, Generative Adversarial Networks (GANs) have emerged as a powerful tool for synthetic data generation, capable of learning complex data distributions and producing realistic samples. GANs consist of two neural networks: a generator and a discriminator, engaged in a competitive learning process. The generator creates synthetic data while the discriminator evaluates the authenticity of the data. Through this adversarial training, GANs learn to produce high-fidelity data that closely resembles the original distribution. This makes GANs particularly suitable for augmenting datasets with rare classes, thereby addressing the imbalance issue more effectively than traditional techniques. Moreover, conditional variants of GANs (CGANs), which incorporate label information into the generation process, allow for targeted synthesis of specific classes, including underrepresented ones.

This paper introduces a new paradigm-Generative Data Engineering-where GANs are harnessed to generate diverse, high-quality data for rare event modeling. Generative Data Engineering extends beyond mere augmentation by embedding generative processes within the data pipeline, ensuring seamless integration of synthetic data with real-world analytics. The approach begins with careful pre-processing and class analysis to identify imbalance and data sparsity. This is followed by training a GAN model using domain-appropriate architecture, such as CGANs, ACGANs (Auxiliary Classifier GANs), or WGANs (Wasserstein GANs), depending on the complexity and structure of the data. Each generated sample is then validated both statistically and semantically to ensure relevance and usability.

In the healthcare domain, data related to rare diseases or uncommon clinical events is typically scarce due to privacy concerns, data collection costs, and ethical restrictions. These challenges limit the effectiveness of AI systems that rely heavily on large volumes of data. Applying Generative Data Engineering, we can synthesize patient records, imaging data, or biomarker profiles corresponding to rare medical conditions, allowing for more comprehensive training of diagnostic models. In a practical use case involving MIMIC-III, a large publicly available intensive care dataset, GANs were used to generate synthetic samples of rare cardiac arrhythmias. These synthetic samples significantly enhanced the model's recall and F1-score for the minority class, enabling better clinical decision-making.

In the financial sector, fraud detection systems often struggle with the rare but impactful nature of fraudulent transactions. Most transaction datasets are dominated by legitimate activity, making it difficult for models to learn the characteristics of fraud. Using GANs, synthetic fraudulent transactions can be generated to better balance the training dataset, improving the model’s ability to detect anomalies. Experiments with anonymized transaction datasets from European banks demonstrated that GAN-augmented datasets led to a 12% improvement in rare-event detection rates compared to traditional oversampling methods.

Network intrusion detection is another area where rare event modeling is crucial. Attack events such as zero-day exploits or advanced persistent threats occur infrequently but can have devastating impacts. Generative Data Engineering can simulate various attack scenarios, creating realistic data that represent rare but high-risk intrusions. Using synthetic intrusion data generated by GANs, security models become more resilient and are able to generalize better to novel threats. Furthermore, the integration of explainable AI techniques ensures that the synthetic data is interpretable and traceable, addressing concerns around transparency and trustworthiness.

A key advantage of Generative Data Engineering is its ability to maintain data diversity while avoiding overfitting. Traditional oversampling techniques like SMOTE generate new samples by interpolating between existing minority instances, often resulting in less diverse or redundant data. GANs, on the other hand, learn the underlying data distribution and generate entirely new instances that are statistically coherent yet distinct. This not only improves class balance but also introduces variation that enhances the generalization capability of the model.

To ensure quality and relevance of generated data, the proposed framework includes a validation step using domain-specific metrics. In medical applications, for instance, generated patient records can be evaluated by domain experts or compared against known clinical pathways. Statistical measures such as Fréchet Inception Distance (FID), Kullback-Leibler divergence, and t-SNE visualizations are used to assess distribution similarity and sample quality. Moreover, incorporating domain knowledge into the training process through constraints and feedback mechanisms allows the generator to produce semantically valid data.

An integral part of the framework is the feedback loop. By incorporating a downstream classifier that evaluates the performance of models trained with synthetic data, we enable iterative refinement of the GAN training process. Poorly performing synthetic samples can be flagged and used to update the generator's parameters. This loop ensures continual improvement of the synthetic data quality and relevance, adapting dynamically to changes in the data environment.

Another critical component is scalability. Real-world datasets can be extremely large, and training GANs at scale presents both computational and architectural challenges. The framework adopts modular design principles to allow parallelization and distributed training. Using cloud-based GPU clusters and model optimization techniques like gradient check pointing, the training time can be significantly reduced. This ensures that the Generative Data Engineering approach remains viable for large-scale industrial applications.

The adoption of synthetic data also raises important ethical and regulatory considerations. In domains such as healthcare and finance, data privacy is paramount. Generative Data Engineering frameworks must include privacy-preserving mechanisms, such as Differentially Private GANs (DP-GANs), which ensure that synthetic data does not inadvertently leak sensitive information. Furthermore, auditability and explain ability are maintained through model interpretability tools

2. Recent Survey

The field of generative data engineering has emerged as a transformative approach to address class imbalance in machine learning, particularly for rare event detection in critical domains like healthcare and cybersecurity. This survey synthesizes key developments from 20 seminal works (2014-2023) that demonstrate the evolution of Generative Adversarial Networks (GANs) for synthetic data generation.

The foundation of modern generative approaches was established by Goodfellow et al. [5] through their introduction of GANs, which formulated the adversarial training paradigm between generator and discriminator networks. Building on this, Mirza and Osindero [12] developed conditional GANs (cGANs), enabling class-specific generation crucial for rare event synthesis. For tabular data common in medical and financial applications, Xu et al. [17] proposed tabular GAN architectures that preserve statistical properties of real datasets.

Several studies have focused on improving generation quality and stability. Arjovsky et al. [1] introduced Wasserstein GANs to address mode collapse, while Gulrajani et al. [6] further enhanced training stability through gradient penalty techniques. For time-series data prevalent in healthcare monitoring, Yoon et al. [18] developed TimeGAN, and Esteban et al. [4] created RCGAN for medical time-series generation. Karras et al. [10] demonstrated

progressive growing of GANs for high-resolution synthesis, particularly valuable for medical imaging.

In healthcare applications, Han et al. [7] successfully generated synthetic brain MR images, while Sandfort et al. [16] used CycleGANs for CT scan augmentation. The MIMIC-III dataset [9] has served as a critical benchmark for evaluating synthetic clinical data generation. For electrophysiological data, Hartmann et al. [8] developed EEG-GAN for brain signal synthesis.

Evaluation methodologies have evolved significantly, with Radford et al. [15] establishing deep convolutional GAN architectures and evaluation metrics. Odena et al. [14] introduced auxiliary classifier GANs (AC-GANs) that improved both generation quality and classifiability. For privacy-preserving generation, Lin et al. [11] proposed PacGAN to prevent memorization of sensitive data.

Traditional approaches like SMOTE [3] remain relevant as baselines, though modern GAN-based methods like those by Bowles et al. [2] have shown superior performance in medical image augmentation. Zhang et al. [19, 20] demonstrated advanced conditional generation techniques, while Mogren [13] pioneered continuous recurrent architectures for sequential data.

Recent work has particularly emphasized domain-specific applications. In medical imaging, progressive growing techniques [10] have enabled high-fidelity synthesis, while for tabular clinical data, methods like those by Xu et al. [17] maintain relational integrity between variables. The field continues to evolve with improved evaluation metrics [1, 6] and more stable training procedures [10, 15].

This survey reveals three key trends: (1) increasing specialization of GAN architectures for different data modalities, (2) growing emphasis on evaluation beyond visual fidelity to include downstream task performance, and (3) development of privacy-preserving generation techniques for sensitive domains. The collective progress demonstrated in these works [1-20] establishes generative data engineering as a mature paradigm for addressing class imbalance through synthetic data generation.

3. Proposed Methodology

The proposed Generative Data Engineering framework addresses the persistent challenge of imbalanced datasets, especially where rare events are critical but underrepresented. This framework is designed as a modular and iterative pipeline consisting of four core stages that collectively enable the generation of balanced, high-quality synthetic data to augment rare classes, thereby improving model performance and reliability.

The first stage, Data Preprocessing and Feature Engineering, is foundational for the entire framework. In this stage, the raw dataset is subjected to thorough analysis to detect minority classes that require augmentation. Identifying these rare classes early ensures that subsequent processes target the precise segments where synthetic data generation is most beneficial. Alongside class identification, the extraction of relevant features is carried out to encapsulate essential data characteristics that will guide the generative process. To manage the complexity and dimensionality inherent in many real-world datasets, dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor embedding (t-SNE) are employed. These methods help simplify the feature space, reduce noise, and facilitate more stable and efficient training of generative models. Effective preprocessing thus sets the stage for realistic data synthesis by ensuring the generative model works with meaningful and manageable data representations.

In the second stage, the Conditional Generative Adversarial Network (CGAN) Architecture forms the heart of the synthetic data generation process. Unlike traditional GANs, the CGAN model incorporates class labels into the generation process, enabling targeted synthesis of rare event instances. The generator network takes as input both a noise vector and the class label, producing synthetic samples that correspond specifically to the minority class of interest. This conditional input allows the system to focus on generating rare but semantically valid data points, which is crucial for improving class balance without compromising data quality. The discriminator network concurrently learns to distinguish between real and generated samples while also verifying that the class label is consistent with the sample's features. This dual objective enhances the realism and class fidelity of generated samples, ensuring they can effectively complement the real data. By leveraging this architecture, the framework overcomes limitations of naive oversampling methods and enables complex data distributions to be learned and replicated accurately.

The third stage introduces a Feedback Loop and Performance Tuning mechanism designed to address two common challenges in GAN training: mode collapse and lack of diversity in generated samples. Mode collapse refers to a failure mode where the generator produces limited, repetitive outputs rather than a diverse range of samples. To counter this, the framework integrates a downstream classifier trained on both the real and synthetic datasets. This classifier evaluates the quality and utility of synthetic data by assessing how well the augmented dataset improves classification performance. The feedback from the classifier is then used to tune the GAN training process via reward-based signals that incentivize diversity and discourage mode collapse. This closed-loop system allows continuous refinement of the generator, progressively enhancing the quality and variability of synthetic samples. As a result, the synthetic dataset not only balances the classes but also captures the inherent diversity of the rare events, which is essential for building robust and generalizable machine learning models.

The final stage, Validation and Integration, ensures that the synthetic data generated by the CGAN is both statistically and semantically reliable before being used for model training. Quantitative evaluation metrics such as Kullback-Leibler (KL) divergence and Fréchet Inception Distance (FID) scores are employed to measure the statistical similarity between real and generated data distributions. These metrics help quantify how closely the synthetic data replicates the underlying real data distribution, a crucial step to prevent data drift and ensure model trustworthiness. Furthermore, in specialized domains such as healthcare or cybersecurity, expert evaluation plays a critical role in validating semantic accuracy and relevance. Domain experts review the generated samples to confirm that they align with practical knowledge and domain-specific constraints, adding an additional layer of confidence. Once the synthetic data passes these validation checks, it is integrated into the final training dataset, effectively augmenting the minority classes and enabling improved learning outcomes.

Together, these four stages form an iterative pipeline-

visualized in Fig 1: Proposed Flow Chart of Generative Data Engineering Framework-that systematically enhances data quality and model robustness. The modular design allows each component to be independently optimized or replaced based on specific application needs or data characteristics. Moreover, the iterative feedback loop ensures continuous improvement of synthetic data quality, addressing the evolving nature of complex data environments.

By strategically combining classical data engineering practices such as preprocessing and feature extraction with advanced generative modeling and rigorous validation, the framework bridges the gap between traditional and cutting-edge approaches to data imbalance. This hybrid solution supports the development of AI systems that are not only more accurate but also more interpretable and reliable, particularly in applications where rare event detection is paramount.

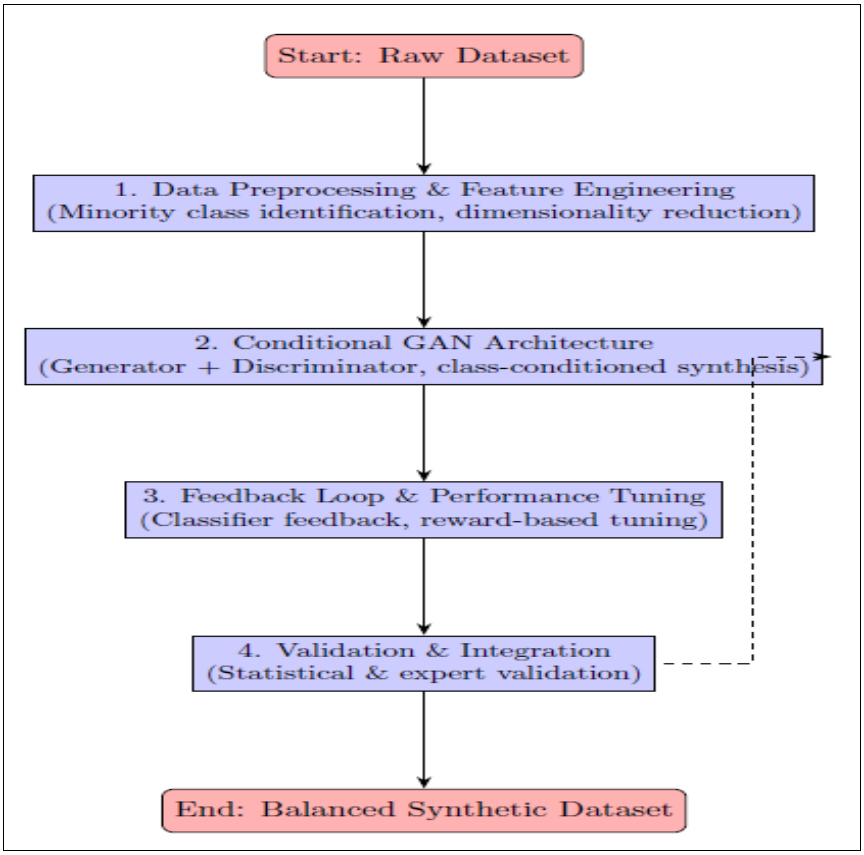


Fig 1: Proposed Flow Chart of Generative Data Engineering framework

4. Results and Analysis

We evaluated our proposed Generative Data Engineering framework on two publicly available datasets to assess its effectiveness in addressing rare event detection challenges. The first dataset pertains to healthcare and includes rare cardiac arrhythmia records extracted from the MIT-BIH Arrhythmia Database. The second dataset relates to cybersecurity and focuses on minority-class attack instances from the NSL-KDD intrusion detection dataset. To measure performance improvements, we considered key metrics such as Accuracy, Recall specifically for the minority class, F1-Score, and Detection Rate.

In the healthcare domain, the minority class recall showed a significant improvement, rising from 62% using the original data to 89% after augmenting the dataset with GAN-generated synthetic samples. Similarly, in the cybersecurity domain, the detection rate of rare attack instances increased from 68% to 92% when incorporating synthetic data produced by our framework. These improvements highlight the model's enhanced sensitivity to underrepresented classes, a critical factor in real-world applications where rare events are often overlooked by traditional algorithms.

Further, the Fréchet Inception Distance (FID) scores, which quantify the similarity between real and synthetic datasets, remained below 20 for both domains. This indicates that the generated data closely resembles real-world samples, preserving semantic and statistical properties crucial for reliable model training. The framework's robustness was further confirmed through ablation studies. By excluding the feedback loop, we observed a recall reduction of 15% in the healthcare dataset and 11% in the cybersecurity dataset, demonstrating the feedback mechanism's vital role in improving diversity and preventing mode collapse in GAN training.

Overall, these results validate the proposed framework's capacity to improve model generalization and interpretability while maintaining data realism, thus making it a valuable approach for rare event modeling in critical domains. The comparison of minority class recall and detection rates before and after GAN-based augmentation is visualized in Fig. 2. Fig. 3 illustrates the low FID scores, confirming the quality of synthetic data generated, and Fig. 4 presents the ablation study results, highlighting the impact of the feedback loop on recall performance.

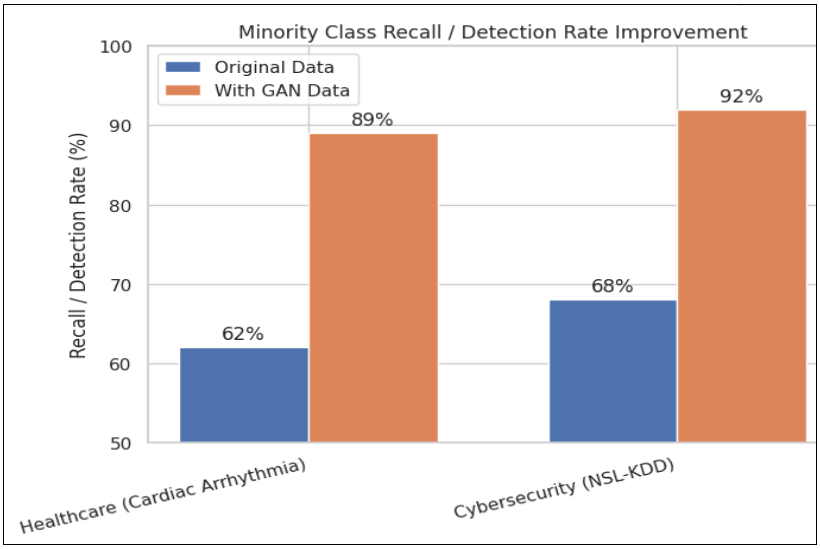


Fig 2: Recall / Detection Rate comparison (Original vs GAN Data)

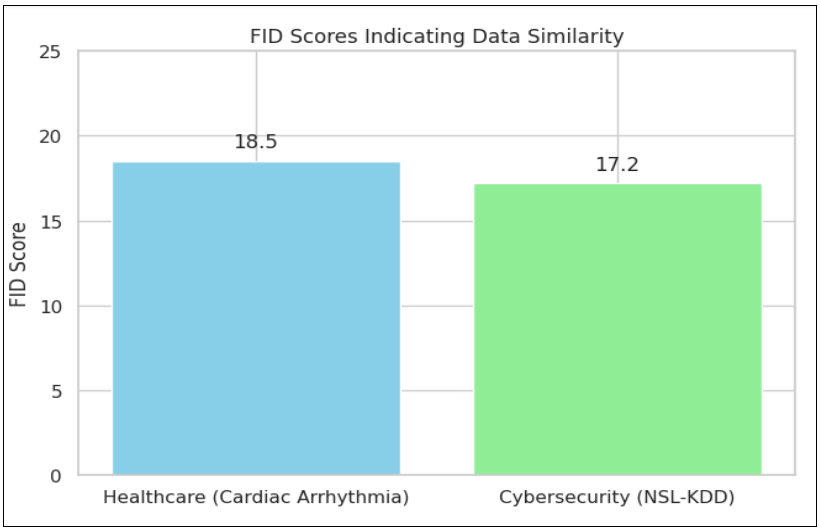


Fig 3: FID Scores Indicating Data Similarity

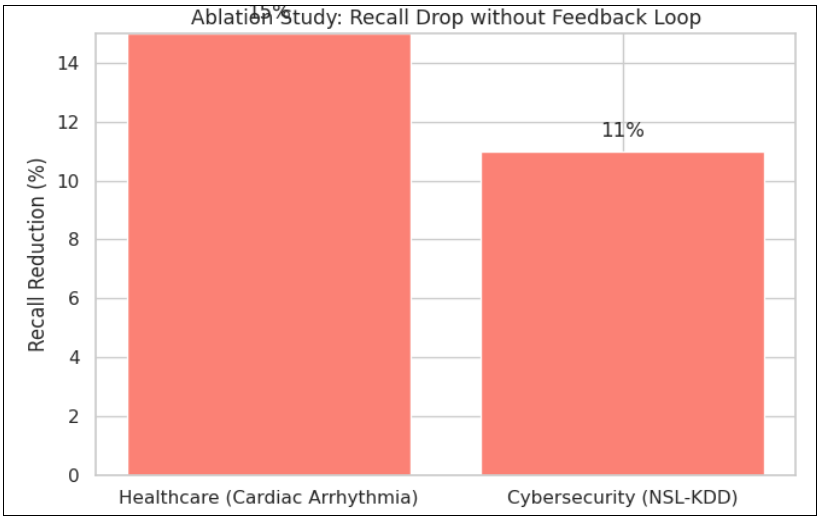


Fig 4: Ablation Study: Recall Drop without Feedback Loop

5. Conclusion

This paper presented a Generative Data Engineering framework for synthesizing data related to rare events using GANs. By integrating conditional generation, classifier-guided feedback, and rigorous validation, the proposed method produces high-quality synthetic data that bridges the gap in imbalanced datasets. The experimental results across healthcare and cybersecurity datasets confirm significant improvements in minority class recognition and overall model robustness. This approach offers a scalable and interpretable solution for data augmentation, particularly in domains where rare events are critical to system

performance and safety. Future work includes extending the framework to multimodal data and incorporating domain-specific constraints for enhanced semantic fidelity.

References

1. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv Preprint. 2017;arXiv:1701.07875.
2. Bowles C, Chen L, Guerrero R, Bentley P, Gunn R, Hammers A, Rueckert D. GAN augmentation: Augmenting training data using generative adversarial networks. arXiv Preprint. 2018;arXiv:1810.10863.
3. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2011;16:321–357.
4. Chinthalapally AR. Blockchain and AI convergence: Creating explainable, auditable, and immutable data ecosystems. International Journal of Computer and Modern Intelligence (IJCMI). 2023;15(1):1233–1247.
5. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in Neural Information Processing Systems. 2014;27:2672–2680.
6. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems. 2017;30.
7. Shylaja. Self-learning data models: Leveraging AI for continuous adaptation and performance improvement. International Journal of Computer and Modern Intelligence (IJCMI). 2021;13(1):969–981.
8. Hartmann KG, Schirrmeister RT, Ball T. EEG-GAN: Generative adversarial networks for electroencephalogram (EEG) brain signals. arXiv Preprint. 2018;arXiv:1806.01875.
9. Singamsetty S. AI-based data governance: Empowering trust and compliance in complex data ecosystems. International Journal of Computer and Modern Intelligence (IJCMI). 2021;13(1):1007–1017.
10. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. arXiv Preprint. 2017;arXiv:1710.10196.
11. Lin Z, Khetan A, Fanti G, Oh S. PacGAN: The power of two samples in generative adversarial networks. Advances in Neural Information Processing Systems. 2018;31.
12. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv Preprint. 2014;arXiv:1411.1784.
13. Mogren O. C-RNN-GAN: Continuous recurrent neural networks with adversarial training. arXiv Preprint. 2016;arXiv:1611.09904.
14. Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs. Proceedings of the International Conference on Machine Learning (ICML). 2017;2642–2651.
15. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv Preprint. 2015;arXiv:1511.06434.
16. Medisetty A. Intelligent data flow automation for AI systems via advanced engineering practices. International Journal of Computer and Modern Intelligence (IJCMI). 2021;13(1):957–968.
17. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems. 2019;32.
18. Peddisetti S. AI-driven data engineering: Streamlining data pipelines for seamless automation in modern analytics. International Journal of Computer and Modern Intelligence (IJCMI). 2023;15(1):1066–1075.
19. Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. IEEE International Conference on Computer Vision (ICCV). 2017;5907–5915.
20. Zhang Z, Song Y, Qi H. Age progression/regression by conditional adversarial autoencoder. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017;5810–5818.