Synthetic Data Generation with Generative AI for Low-Resource Predictive Modeling

Emile C. Crawford
Department of Computer Science, University of North Texas, Denton, TX, USA.
emile.crawford778@unt.edu

Karan L. Sharma
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
karans@binghamton.edu

Benjamin Lawrence
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
benjaminl@colostate.edu

Abstract

The increasing demand for predictive models in domains where labeled data are scarce has motivated the exploration of synthetic data generation using generative artificial intelligence. This paper presents a comprehensive systems-level analysis of the architectures, trade-offs, and governance challenges inherent in deploying generative models for low-resource predictive modeling. We examine the structural properties of generative adversarial networks, variational autoencoders, and diffusion models as they relate to data fidelity, diversity, and privacy preservation. The discussion extends to the infrastructure required for training such models on limited real data, including transfer learning, self-supervised pretraining, and federated setups. Critical tensions between statistical utility and fairness are analyzed, particularly the risk of amplifying biases present in small real-world samples. Policy implications, including regulatory frameworks for synthetic data provenance and accountability, are explored through cross-domain case illustrations in healthcare, finance, and natural language processing. The paper concludes with forward-looking perspectives on sustainable deployment, robustness evaluation, and the need for standardized benchmarks. Our findings suggest that while generative AI offers a powerful pathway to overcome data scarcity, its success depends on careful architectural choices, rigorous validation protocols, and governance structures that ensure equitable outcomes across populations.

Keywords

synthetic data, generative AI, low-resource predictive modeling, data augmentation, fairness, robustness, governance, infrastructure.

1. Introduction

Predictive modeling has become a cornerstone of decision-making across healthcare, finance, autonomous systems, and policy analysis. However, many high-impact applications suffer from a fundamental data scarcity problem: labeled examples are expensive to obtain, privacy-sensitive, or simply unavailable in sufficient quantity to train robust models. This low-resource regime challenges traditional supervised learning paradigms and has spurred interest in synthetic data generation as a means to augment limited real datasets. Generative artificial intelligence, encompassing models such as generative adversarial networks (GANs),

variational autoencoders (VAEs), and more recently diffusion models, offers the ability to produce new samples that mimic the statistical properties of the original data without directly exposing individual records [1], [2], [3]. The promise is that synthetic data can fill gaps, improve model generalization, and enable predictive analytics in contexts where data collection is infeasible. Yet the adoption of synthetic data introduces systemic complexities that extend far beyond model architecture. The reliability of downstream predictions depends critically on how well synthetic data capture the underlying distribution, whether they preserve rare but important patterns, and how they interact with the biases inherent in the original sample. Furthermore, the deployment of generative models in low-resource settings imposes constraints on computational infrastructure, validation methods, and governance structures that are often underappreciated in the literature. This paper adopts a systems-level perspective to examine the interplay between generative AI design choices and the practical demands of low-resource predictive modeling. We explore structural trade-offs between data fidelity and privacy, the architectural considerations that determine scalability, and the policy implications of relying on artificially generated observations for consequential decisions. By situating these technical discussions within broader socio-technical contexts, we aim to provide a framework that researchers and practitioners can use to evaluate the suitability of synthetic data for their specific low-resource use cases.

2. Background and Related Work

The field of synthetic data generation has evolved rapidly over the past decade, driven by advances in deep generative models. Early efforts using simple statistical imputation and oversampling techniques, such as SMOTE, provided modest improvements but struggled to capture complex multivariate dependencies [4]. The introduction of generative adversarial networks by Goodfellow et al. marked a paradigm shift, enabling the generation of high-dimensional data such as images and time series through an adversarial training process [5]. Subsequent refinements, including Wasserstein GANs and conditional GANs, improved training stability and allowed for class-conditional synthesis [6], [7]. Variational autoencoders offered an alternative framework based on probabilistic inference, producing smoother latent spaces that facilitate interpolation and controlled generation [8]. More recently, denoising diffusion probabilistic models have achieved state-of-the-art sample quality in image generation, with applications extending to tabular and sequence data [9]. In parallel, research on data augmentation for low-resource natural language processing has leveraged back-translation, word-level substitutions, and generative models to create paraphrases and synthetic text [10]. Transfer learning, where a generative model is pretrained on a large corpus and fine-tuned on a small target dataset, has become a common strategy to mitigate data scarcity [11]. Despite these advances, the literature has primarily focused on model performance metrics such as inception score, Fréchet inception distance, and downstream accuracy, while less attention has been paid to the systemic implications of synthetic data use. Issues of fairness, privacy, and domain shift are often treated as secondary considerations. This paper builds on existing work by integrating these dimensions into a unified systems analysis, emphasizing how architectural decisions affect not only predictive accuracy but also the robustness and ethical defensibility of deployed models.

3. Architectural Considerations for Synthetic Data Generation

The choice of generative architecture profoundly influences the quality and utility of synthetic data in low-resource settings. Generative adversarial networks consist of a generator and discriminator that compete in a minimax game, forcing the generator to produce samples that

are indistinguishable from real ones. In practice, GANs are notoriously difficult to train, particularly when the real dataset is small, because the discriminator can quickly memorize the limited examples and the generator may suffer from mode collapse [12]. Conditional GANs mitigate this to some extent by providing auxiliary information, but they still require careful hyperparameter tuning and architectural regularisation. Variational autoencoders offer a more stable training objective based on evidence lower bound optimization, but they tend to produce blurrier samples that may lack fine-grained detail. For low-resource predictive modeling, this trade-off between fidelity and stability is critical: if the synthetic data are too noisy or unrealistic, downstream classifiers may learn spurious correlations or fail to generalize. Diffusion models, which generate data by iteratively denoising a random signal, have demonstrated remarkable fidelity and diversity, yet they require substantial computational resources for both training and sampling. In resource-constrained environments, this cost may be prohibitive. An alternative paradigm is to use transfer learning by initializing the generative model with weights pretrained on a large, publicly available dataset. For example, a GAN pretrained on ImageNet can be fine-tuned on a small medical imaging dataset, leveraging the learned feature hierarchies to produce plausible synthetic radiographs [13]. Similarly, large language models pretrained on massive text corpora can be fine-tuned for synthetic text generation in specialized low-resource domains such as legal or clinical documents. However, transfer learning introduces its own challenges: the pretrained distribution may not align well with the target domain, leading to domain shift and potential biases. Domain adaptation techniques, including adversarial alignment and feature normalization, can help, but they add complexity to the overall pipeline. The architectural decision must therefore balance the need for high-fidelity synthetic data, the constraints of limited training examples, and the computational budget available. In many low-resource scenarios, a hybrid approach that combines pretrained components with lightweight fine-tuning on the target dataset offers a pragmatic solution, though rigorous validation of the synthetic data's statistical properties remains essential.

4. Structural Trade-offs in Data Fidelity and Utility

A central tension in synthetic data generation is the trade-off between fidelity—how closely synthetic samples match the real distribution—and utility, defined as the improvement in downstream predictive model performance. High-fidelity synthetic data that are nearly indistinguishable from real observations can inadvertently propagate the same noise, outliers, and biases present in the original small sample. Conversely, synthetic data that are intentionally smoothed or privatized may lose important rare patterns, such as long-tail disease subtypes or minority class features, thereby reducing model robustness for underrepresented groups. This trade-off is not merely statistical but structural, because the generative model's capacity to capture the true data manifold depends on the interplay of architecture, training data size, and regularization. For instance, a GAN trained on a very small dataset may overfit by memorizing the real samples, producing synthetic copies that offer no new information for downstream learning. Alternatively, a VAE with a strong prior may average out variations, generating samples that lie in the center of the data distribution but lack the diversity needed to improve classifier decision boundaries. Privacy-preserving techniques, such as differentially private training of generative models, further degrade fidelity in exchange for formal guarantees against re-identification [14]. In low-resource contexts, the privacy-utility trade-off is especially acute because adding noise to protect a small number of individuals can distort the entire synthetic distribution. The relevant literature has proposed metrics to quantify these trade-offs, including membership inference

risk, distance-based fidelity measures, and downstream accuracy curves. However, these metrics are often computed on held-out real data that are themselves scarce, making evaluation challenging. A robust approach involves using a separate validation set, cross-validation, or synthetic data for evaluation only when combined with domain expertise. Ultimately, the decision to prioritize fidelity or utility must be guided by the specific predictive task: for risk scoring in rare diseases, preserving minority patterns may be paramount, whereas for aggregate trend analysis, smoothed synthetic data may suffice.

5. Governance, Fairness, and Policy Implications

Deploying synthetic data in predictive modeling for high-stakes decisions raises significant governance and fairness concerns that require careful policy attention. Generative models trained on real data that contain historical biases—such as racial disparities in medical records or gender imbalances in financial credit data—will likely replicate and even amplify those biases in synthetic samples [15]. When downstream models are trained on such synthetic data, the resulting predictions can perpetuate systemic inequities, undermining the goal of fairness. Moreover, the lack of transparency in deep generative models makes it difficult to audit how biases propagate. One proposed mitigation is to incorporate fairness constraints during generative model training, for example by minimizing correlation between synthetic features and protected attributes or by reweighting the training distribution to ensure balanced representation [16]. However, such interventions require access to protected attribute information, which may be unavailable or itself subject to privacy regulations. Another governance challenge concerns the provenance and accountability of synthetic data. Unlike real data, which are collected through identifiable processes, synthetic records do not correspond to actual individuals, raising questions about liability if a model trained on synthetic data makes a harmful prediction. Regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR) and the proposed Artificial Intelligence Act have begun to address synthetic data, but clear guidelines on validation, documentation, and transparency are still evolving [17]. For instance, synthetic data used in clinical decision support must undergo rigorous evaluation to ensure that predicted outcomes are clinically meaningful and not artifacts of the generative process. Policy makers and institutional review boards should require that synthetic datasets be accompanied by metadata describing the generative model architecture, training conditions, and fidelity assessments. Furthermore, the long-term storage and reusability of synthetic data raise sustainability issues: as real-world populations change, synthetic data become stale and may induce distributional shift. Governance structures must include periodic updates and retirement mechanisms for synthetic datasets. Finally, the development of standardized benchmarks for evaluating synthetic data quality across domains would facilitate regulatory oversight and enable cross-study comparisons.

6. Deployment and Infrastructure Challenges

The practical deployment of generative AI for low-resource predictive modeling requires a robust infrastructure that addresses data storage, computational resources, model lifecycle management, and integration with existing systems. In many low-resource settings, such as rural healthcare clinics or small financial institutions, the available computing hardware may be limited to modest workstations or cloud-based services with restricted budgets. Training large diffusion models or GANs from scratch on such hardware is infeasible, necessitating the use of lightweight architectures, quantization, or distillation techniques. Edge deployment, where synthetic data generation occurs on local devices, introduces additional constraints on

memory and energy consumption. A promising direction is the use of federated learning frameworks that allow multiple low-resource sites to collaboratively train a generative model without sharing raw data [18]. In a federated setup, each site trains a local generator on its own small dataset and periodically shares model updates with a central server, which aggregates them to produce a global generative model. This approach preserves data privacy and can improve synthetic data quality by leveraging diverse samples across sites. However, federated training introduces communication overhead, synchronization challenges, and vulnerability to adversarial attacks. Infrastructure planning must therefore consider the trade-off between model quality and operational cost. Another deployment challenge is the integration of synthetic data into existing predictive modeling pipelines. Standard machine learning workflows assume that training and test data come from the same distribution; when synthetic data are used for augmentation, the distributional mismatch can degrade calibration and uncertainty estimates. Monitoring systems that detect distributional drift between synthetic and real data are essential for maintaining model reliability over time. Robustness testing should include adversarial scenarios where the generative model's biases are exploited. For example, if a synthetic data generator only produces images under controlled lighting conditions, a downstream object detector trained on such data may fail in real-world variable lighting. Comprehensive stress testing, including sensitivity analysis and out-of-distribution detection, should be part of the deployment protocol. Additionally, the computational cost of generating synthetic data on demand must be accounted for in latency-sensitive applications such as real-time fraud detection. Caching strategies and precomputed synthetic datasets can mitigate this, but they reduce flexibility. Ultimately, a successful deployment requires a system-level perspective that interconnects the generative model, the predictive model, the evaluation framework, and the operational environment.

7. Case Illustrations and Cross-Domain Comparisons

To illustrate the systemic considerations discussed above, we examine three domains where synthetic data generation has been applied to low-resource predictive modeling: healthcare, finance, and natural language processing. In healthcare, synthetic electronic health records (EHRs) have been generated using GANs and VAEs to augment datasets for predicting patient readmission, disease progression, and treatment response [19]. A key challenge is that real EHRs contain complex temporal dependencies, missing values, and high-dimensional codes. Early attempts produced synthetic records that did not preserve the temporal correlations, leading to poor downstream performance. Later work used recurrent generative models and attention mechanisms to capture sequence structure. Privacy regulations such as HIPAA in the United States restrict the sharing of real patient data, making synthetic data an attractive alternative for multi-institutional studies. However, biases in the real data—such as underrepresentation of minority ethnic groups—can be magnified in synthetic records, potentially leading to models that perform poorly for those groups. In finance, synthetic transaction data have been used to train fraud detection models where real fraudulent transactions are rare. Generative models must preserve the long-tail distribution of fraud patterns while avoiding the generation of synthetic frauds that are too easy to detect (and thus provide no learning signal) or too similar to real frauds (risking re-identification). Additionally, financial regulations require explainability, and synthetic data generators are often black boxes, complicating compliance. Some approaches have combined GANs with rule-based constraints to ensure that synthetic transactions adhere to domain-specific business logic. In natural language processing, synthetic text generation has been used for low-resource sentiment analysis, named entity recognition, and question answering. Large language models

fine-tuned on small domain-specific corpora can produce coherent synthetic sentences, but they may hallucinate facts or introduce stylistic artifacts that mislead downstream classifiers. For languages with little digital data, such as many indigenous languages, synthetic text generation holds promise but requires careful alignment with linguistic structures to avoid cultural misrepresentation. Cross-domain comparison reveals that while the core generative architectures are transferable, the specific trade-offs differ: in healthcare, privacy and clinical validity are paramount; in finance, regulatory compliance and explainability dominate; in NLP, semantic fidelity and diversity are critical. These examples underscore the need for domain-adapted evaluation frameworks and governance mechanisms.

8. Future Directions and Sustainability

The field of synthetic data generation for low-resource predictive modeling is still maturing, and several future directions promise to address current limitations. One important avenue is the development of foundation models specifically designed for low-resource settings, such as small-scale generative models that can be efficiently trained on limited data while maintaining high fidelity. Research on neuromorphic computing and spiking neural networks may offer energy-efficient alternatives for edge-based generation. Another direction involves the integration of causal generative models that explicitly model the underlying data generation process rather than merely the observed distribution. Causal synthetic data could improve out-of-distribution generalization and enable counterfactual reasoning, which is particularly valuable in policy evaluation and personalized medicine [20]. Sustainability concerns also demand attention: the carbon footprint of training large generative models is substantial, and for low-resource applications, lighter models that require less energy are preferable. Techniques such as knowledge distillation, pruning, and early stopping can reduce computational costs without sacrificing too much quality. Furthermore, the development of standardized synthetic data benchmarks that include fairness, privacy, and robustness metrics would accelerate progress and facilitate cross-study comparisons. Currently, most benchmarks focus on image generation; tabular and sequential data benchmarks are needed. In terms of governance, future regulatory frameworks should mandate that synthetic data used in consequential applications be accompanied by a "nutrition label" that discloses the generative model, training data composition, known biases, and fidelity metrics. Finally, interdisciplinary collaboration among computer scientists, domain experts, ethicists, and policymakers will be essential to ensure that synthetic data generation serves equitable and sustainable ends. As generative AI continues to evolve, the tension between its potential and its risks will only intensify, making systems-level analysis a crucial component of responsible innovation.

9. Conclusion

This paper has provided a comprehensive systems-level examination of synthetic data generation using generative AI for predictive modeling in low-resource contexts. We have argued that the effectiveness of synthetic data depends not only on the capabilities of generative models but also on a constellation of architectural decisions, infrastructural constraints, and governance structures. The trade-off between data fidelity and utility is central, and its resolution varies by domain and intended application. Privacy, fairness, and robustness must be explicitly engineered into the generative pipeline rather than treated as afterthoughts. Deployment in real-world settings requires careful infrastructure planning, monitoring, and validation protocols that account for distributional shift and the limitations of small real samples. Cross-domain case illustrations highlight the importance of domain-

specific evaluation and adaptation. As synthetic data become more pervasive, the need for standardized benchmarks, transparent documentation, and regulatory oversight will grow. Future work should focus on causal generative models, sustainable architectures, and interdisciplinary governance frameworks. Ultimately, synthetic data generation is a powerful tool, but its use in low-resource predictive modeling demands rigorous systems thinking to ensure that the resulting predictions are accurate, fair, and trustworthy.

References

1. S. He, C. Li, and J. Wang, "Generative adversarial networks for synthetic data generation: A survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 10, pp. 5235–5254, 2022.
2. D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in Proceedings of the International Conference on Learning Representations (ICLR), 2014.
3. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 6840–6851.
4. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
5. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems, vol. 27, 2014.
6. M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in Proceedings of the International Conference on Machine Learning (ICML), 2017, pp. 214–223.
7. M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014.
8. C. Doersch, "Tutorial on variational autoencoders," arXiv preprint arXiv:1606.05908, 2016.
9. P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 8780–8794.
10. T. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 86–96.
11. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in Advances in Neural Information Processing Systems, vol. 27, 2014.
12. I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," arXiv preprint arXiv:1701.00160, 2016.
13. H. Choi, S. Kim, and J. Lee, "Medical image synthesis using generative adversarial networks: A systematic review," IEEE Reviews in Biomedical Engineering, vol. 15, pp. 169–186, 2022.
14. M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proceedings of the ACM Conference on Computer and Communications Security (CCS), 2016, pp. 308–318.

15. R. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1–35, 2021.

16. S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning. MIT Press, 2019.

17. European Commission, "Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," COM(2021) 206 final, 2021.

18. Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 2, pp. 1–19, 2019.

19. E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in Proceedings of the Machine Learning for Healthcare Conference, 2016, pp. 301–318.

20. B. Schölkopf, "Causality for machine learning," in Probabilistic and Causal Inference: The Works of Judea Pearl, 2022, pp. 765–804.