Optimizing Synthetic Data from Scarcity: Towards Meaningful Data Generation in High-Dimensional Low-Sample Size Domains

Danilo Danese[1,*]

[1]Politecnico di Bari, Via E. Orabona, 4, 70126 Bari, Italy

Abstract

Deep learning has revolutionized artificial intelligence by enabling the extraction of intricate representations from large datasets. Generative models have emerged as powerful tools for data synthesis by mimicking the distributions of training data. Despite their advancements, these models encounter critical concerns, including biases, privacy risks, and the authenticity of the generated data. These challenges underscore the necessity for incorporating fairness, expert insights, and comprehensive evaluations into their development. Applications like medical imaging pose challenges due to scarce high-quality data and demanding requirements for condition-specific synthesis. Furthermore, the management of high-dimensional low-sample size (HDLSS) data accentuates the demand for sophisticated representation learning techniques, enabling the generation of effective synthetic data from limited clinical datasets. The complexity of longitudinal medical data, characterized by intricate temporal correlations, further challenges existing methodologies, revealing their limitations. In the light of above, my doctorate research path intends to focus on two main objectives: (i) employ cutting-edge techniques to advance beyond current state-of-the-art in data synthesis, and (ii) bridge the gap between privacy, fairness and generating meaningful synthetic data leveraging on XAI and HCI for further robustness.

Keywords

Synthetic Data Augmentation, Sensitive Domains, Generative Models, XAI, HCI

1. Introduction

Deep learning has revolutionized artificial intelligence by enabling the extraction of intricate patterns from vast datasets. However, in many fields, including healthcare, the scarcity of labeled data remains a significant bottleneck despite the increasing availability of large datasets. This challenge is particularly acute in healthcare due to the limited availability of patient cohorts and the high dimensionality of data, such as neuroimaging with millions of voxels. Traditional statistical analyses may be unreliable [1] due to the sparse representation of the population. While algorithms based on deep learning frameworks show impressive performance, their effectiveness relies heavily on the availability of training samples, often requiring large datasets to avoid overfitting and ensure statistically meaningful results [2]. Acquiring high-quality reference standards for labeling demands substantial investments in time, financial

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.

danilo.danese@poliba.it (D. Danese)

0009-0000-5203-1229 (D. Danese)

resources, and human expertise [3]. Moreover, class imbalance is prevalent in manually labeled healthcare datasets, with certain categories underrepresented. To address the limited availability of medical training data, researchers have proposed various data augmentation (DA) techniques to synthetically generate additional examples. However, these methods have limitations in capturing the full diversity and complexity of real-world clinical data. DA offers a promising approach to expand datasets by generating synthetic, labeled samples. This approach employs a variety of techniques such as the injection of Gaussian noise, cropping, flipping, and padding to generate new samples that retain the original image's label, thereby preserving the semantic meaning of the data [4]. However, the effectiveness of DA depends heavily on the original data, as transformations beneficial for one dataset may introduce bias or be ineffective for another. The susceptibility to misleading transformations is exemplified by rotating a "6" to resemble a "9". This issue is pertinent in medical domains, where the complexity of medical images, including diverse anatomy, irregular tumor shapes, and occasional anatomical inconsistencies, can render traditional operations ineffective. The limitations of DA can result in the generation of irrelevant or anomalous images, disrupting model performance. Additionally, imbalanced class representation can bias models toward overrepresented categories, necessitating careful evaluation and transformation methods to address class imbalance effectively. The primary aim of DA extends beyond augmenting data volume to faithfully replicate the true data distribution. This involves generating synthetic samples that not only blend indistinguishably with the original data but also encapsulate the complex relationships and patterns inherent within the dataset. Synthetic samples generated through DA must not only preserve the label of their source data but also embody the nuanced statistical characteristics that underpin their authenticity. As a result, it should be difficult to discern synthetic samples from real ones.

Balancing Realism and Variability

While imposing constraints to enhance the realism and clinical validity of synthesized medical images is crucial, it is important to acknowledge the potential trade-off between realism and variability. Overly strict restrictions on the generative models may limit the diversity and representative capacity of the generated samples, introducing bias into the synthetic data. Ensuring a high degree of realism is essential for the effective utilization of synthetic data in medical applications. However, if the generative process is excessively constrained, the models may fail to capture the true underlying distribution of the data, leading to an incomplete or skewed representation of the target domain. This could result in synthetic samples that do not accurately reflect the inherent diversity and variability present in real clinical data, potentially hindering the generalization capabilities of models trained on such data.

To address this challenge, it is crucial to strike a delicate balance between imposing constraints for realism and allowing sufficient flexibility for variability. One approach could involve incorporating domain knowledge and expert feedback through human-computer interaction (HCI) techniques. By leveraging the expertise of medical professionals, the generative process can be guided to prioritize clinically relevant features and constraints, while still allowing for a certain degree of variability within acceptable bounds.

Moreover, the integration of explainable AI (XAI) methodologies can provide insights into the generative models' decision-making processes, enabling a more clear understanding of the factors influencing the generated samples. Through XAI-guided analysis, it may be possible

to identify and adjust specific model components or parameters that contribute to unrealistic artifacts or limited variability, without compromising overall realism.

Additionally, iterative refinement cycles with domain experts could be implemented, wherein generated samples undergo evaluation for realism and diversity, leading to adjustments in the generative process. This iterative approach can help fine-tune the balance between realism and variability, ensuring that the synthetic data accurately reflects the complexities and nuances of real clinical data while maintaining representative diversity.

2. Background Technologies

Image augmentation techniques constitute a fundamental subset of DA methods that manipulate existing image samples through various transformations. These can range from basic operations such as elastic transformations which deform image spatial dimensions through displacement fields, introducing localized shape variations without constraints on parallelism or aspect ratios. However, elastic warping may produce anatomically implausible samples. On the other hand, erasing transformations replace selected image regions with constant intensity values or random noise, while pixel-level techniques adjust attributes like brightness, contrast, saturation, and noise. These augmentations mainly modify original data, potentially limiting generalization. Augmented samples also tend to strongly correlate with originals.

Synthetic generation offers a different approach to overcome the limitations of traditional techniques. Unlike manipulating existing data, these methods create entirely new samples from scratch, potentially introducing greater diversity and complexity. Specialized models, tailored to specific modalities and tasks, further enhance those capabilities. While promising, synthetic techniques require increased computational resources and more complex architectures compared to basic transformations. A crucial challenge remains to ensure the visual fidelity and realism of generated samples, as unrealistic artifacts can negatively impact model performance.

Generative Adversarial Networks (GANs). GANs [5] have showcased their capacity to produce lifelike images, rendering them extensively utilized in medical research [6] and incorporated into various DA assessments [7]. However, despite their efficacy, GANs are not without limitations; challenges include learning instability, convergence difficulties, and susceptibility to mode collapse [8], wherein the generator generates a limited number of samples, and therefore limiting their ability to diversify the data and improve model generalizability. Moreover, previous research [9] has shown that GANs can sometimes "hallucinate" features in generated images, potentially introducing artifacts that mimic or hide real features. Medical image datasets frequently display significant class imbalance, with a bias towards healthy or normal cases. One prevalent augmentation strategy involves integrating synthesized pathological lesions into otherwise healthy images. However, Cohen et al. [9] have identified significant challenges with this method, particularly when employing Cycle Generative Adversarial Networks (CycleGANs) for data translation tasks, whether involving unpaired or paired data. These studies demonstrate a significant limitation: CycleGANs may fail to accurately retain all known and potentially unknown class labels during the translation process.

Variational autoencoders (VAEs). VAEs [10] offer greater output diversity and avoid mode

collapse compared to GANs, but they often produce blurry, low-fidelity images due to minimizing the Kullback-Leibler divergence [11]. While VAEs have advantages over GANs in stability and sample variety, their image quality limitations have restricted their adoption for DA. Nonetheless, Chadebec et al. [12] achieved a noteworthy result with the introduction of a geometry-aware VAE tailored for DA in HDLSS scenarios. By integrating Riemannian geometry into the model, they enhanced the learning process of latent representations, enabling the generation of realistic samples even with sparse data. Their model demonstrated substantial performance gains over a standard VAE in an MRI classification task, achieving an approximate 8% increase in accuracy when trained on a dataset consisting of 50 real and 5,000 synthetic MRIs.

Diffusion Models (DMs). Recent academic literature has notably expanded the use of DMs for image synthesis [13]. DMs achieve high-fidelity sample generation by approximating complex real-world data distributions through a series of simpler distributions progressively 'diffused' together. This process effectively captures the intricacies of the original data, resulting in more realistic and diverse generated samples. This capability is particularly advantageous in image synthesis tasks, as general-purpose images often present a diverse array of textures, colors, and other visual attributes that challenge simpler parametric models. However, despite these strengths, DMs also present certain limitations. Compared to GANs and VAEs, DMs can be computationally demanding and require a significant amount of data for accurate calibration. Moreover, DMs entail prolonged sampling times due to the extensive steps in the reverse diffusion process, posing challenges for real-time applications or scenarios requiring a substantial volume of samples. Consequently, researchers propose solutions to enhance sampling efficiency while maintaining sample quality and diversity. A notable approach is progressive distillation [14], which involves distilling a pre-trained deterministic diffusion sampler with numerous steps into a novel diffusion model requiring fewer sampling steps.

While DMs progress rapidly, their application in medical contexts lags, highlighting a gap between state-of-the-art (SotA) generative models in general domains and those in medical fields. There is a pursuit for computational efficiency in general-purpose generative models. Wuerstchen [15], a novel architecture, demonstrates competitive performance and cost-effectiveness for large-scale DMs. Central to Wuerstchen's innovation is its employment of a latent diffusion technique that acquires a detailed semantic image representation, reducing computational costs while surpassing SotA. Wuerstchen achieves comparable results with less data, doubling inference speed and reducing time and costs.

State Space Models (SSMs). Parallel to DMs, research explores computationally efficient architectures. Mamba [16], spanning modalities like language, audio, and genomics, utilizes efficient Selective State Space Models for sequence processing, demonstrating speed improvements compared to Transformers [17] particularly evident when handling longer sequences. Applying SSMs to vision is challenging due to data characteristics. Vision-specific architectures like Vim [18], with bidirectional Mamba blocks and position embeddings, aim to overcome limitations of self-attention. Another work [19] inspired by SSMs and Vision Transformers models maintains global receptive fields while improving efficiency. Notably, addressing direction-sensitivity in images enables the processing of visual data as ordered sequences. These architectures show promise in various vision tasks, especially at higher resolutions, potentially paving the way for sensitive domains like medical imaging.

3. Exploratory Approach: Integrating XAI and RLHF

Explainable Artificial Intelligence (XAI). XAI has emerged as a critical field in the era of deep learning, aiming to address the opaque and complex nature of modern machine learning models. As these models become increasingly sophisticated and are deployed in sensitive domains, the need for transparency, interpretability, and accountability has become fundamental. XAI techniques aim to clarify how these models make decisions, helping users and researchers understand why they make certain predictions, recognize any biases, and ensure compliance with ethical and regulatory guidelines. In the context of synthetic data generation, XAI plays a pivotal role in ensuring the reliability and trustworthiness of the generated data.

As previously mentioned, synthetic DA is fundamental in fields with scarce, sensitive, or costly real-world data. Among GANs, VAEs, SSMs and DMs, the latter have gained significant attention due to their ability to generate high-fidelity samples by approximating complex real-world data distributions, yet their complex architectures and iterative generation processes pose challenges in terms of interpretability and explainability. Integrating XAI techniques into DMs for synthetic data generation can provide valuable insights into the model's decision-making process, as proposed by Park et al. [20]. By visualizing and interpreting the denoising process, researchers can identify the regions and visual concepts that the model focuses on at each time step, ensuring that the generated data accurately captures the desired features.

Although initially designed for general-purpose images, its application within the medical domain proves advantageous for ensuring accurate representation and generation of relevant anatomical structures and pathological features. An initial strategy for developing forthcoming frameworks may involve the adaptation of tools such as DF-RISE and DF-CAM [20], derived from RISE and Grad-CAM. DF-RISE and DF-CAM are complementary techniques that provide insights into the DM's decision-making process from external and internal perspectives, respectively. DF-RISE reveals the denoising levels and regions of focus, while DF-CAM unveils the specific visual concepts prioritized by the model at each step In DMs. This approach can facilitate the comprehension of which visual concepts (e.g., specific organs, tissues, lesions) are prioritised at different time steps during image synthesis. It can also aid in fine-tuning the model to better capture the desired features during experiments. By visualising the denoising levels using DF-RISE, it is possible to comprehend the semantic and detail levels recovered by the DM during generation. This can help ensure that the model accurately captures both high-level semantic information (e.g. organ structures) and fine-grained details (e.g. lesions, abnormalities). Furthermore, DF-CAM can help visualize and interpret the visual concepts the DM focuses on at each inference step during medical image generation. This can provide insights into the model's decision-making process and assist in identifying potential biases or inconsistencies in the generated images.

Reinforcement Learning from Human Feedback (RLHF). The synergy between XAI and DMs can be further amplified through RLHF [21] to improve the generation process. In the context of DMs, RLHF enables learning from human preferences and feedback, enhancing their ability to generate data that aligns with human expectations [22]. The integration of XAI and RLHF facilitates the acquisition of insights into the decision-making mechanisms of the model and enables active refinement of the generation process to better align with human values.

In this task, RLHF can be used to fine-tune DMs for generating synthetic medical images that align with medical standards and requirements. The process involves training a DM to generate medical images based on specific classes. Human evaluators (e.g. domain experts) then provide feedback on the generated images, indicating which ones better align with the task requirements, such as improving class-image alignment, or refining aesthetic quality. In this work [22] was proposed the D3PO method, an extension of DPO, which directly fine-tunes DMs based on human feedback without requiring a separate reward model. This approach is more direct, cost-effective, and minimizes computational overhead compared to traditional RLHF methods that rely on a reward model and are incompatible with the strict requirement of a domain expert supervision. Physicians can provide valuable feedback on the quality and relevance of medical images, guiding the fine-tuning process of DMs.

4. Research Questions and Methodology

This research work aims to investigate the capabilities and the limits of modern generative models in the context of HDLSS domains. Specifically, the primary objective of this study is to develop a comprehensive framework capable of effectively addressing the fundamental constraints associated with scenarios characterized by HDLSS data such as medical imaging. This framework will focus on systematically addressing challenges related to privacy, fairness, and consistency in discrete segments. To ensure the reliability and robustness of the methodology, the incorporation of Human-Computer Interaction (HCI) techniques can facilitate systematic validation of significant outcomes. By harnessing the expertise of domain specialists, these techniques can guide both the generation of synthetic data and the final output, thereby validating the integrity and credibility of the framework. Moreover, it is imperative to devise a solution that transcends limitations inherent to specific environments by leveraging reproducible datasets and extending its applicability to a diverse array of real-world datasets, as real-world data may exhibit unpredictable or anomalous behavior.

RQ1: What are the current limitations of using modern generative models for sensitive domains?
Modern generative models have immense potential to benefit healthcare, but their adoption remains limited. Key challenges include: ensuring the clinical validity of synthesized images; handling multimodal data; scarce annotated datasets; protecting privacy; explaining outputs; computational efficiency; and robustness to data distribution shifts. A comprehensive examination of these limitations can help identify critical gaps and requirements, informing the development of tailored methodologies, validation procedures, guidelines, and best practices.

Specifically, how to ensure that data from the same patient or source remains entirely within either the training or validation set, preventing any overlap or leakage between the two? Additionally, how to establish appropriate similarity measures that go beyond human visual evaluation to objectively assess whether the synthetic data accurately captures the true underlying distribution of the target domain?

To address this, it is necessary to develop robust evaluation protocols that involve splitting the limited dataset into two distinct subsets (training and validation) in a principled manner, ensuring that data from the same source (e.g., patient) is consistently allocated to either subset.

This strict separation is crucial to maintain the validity of the DA process and subsequent model evaluation, restricting the generation of synthetic data solely to the training subset.

However, rather than relying solely on human visual inspection, which can be subjective and limited, we must establish quantitative similarity measures that can objectively evaluate the extent to which the synthetic data accurately represents the characteristics and nuances present in the validation set. These measures should be designed to assess the ability of models trained on the augmented data to generalize and accurately label or process the validation data, thereby indirectly evaluating the quality and representativeness of the synthetic samples.

Finally, it is essential to acknowledge that in certain scenarios, the available data may be so scarce that confident label prediction becomes impossible without introducing additional real data into the dataset. In such extremely data-limited situations, relying solely on augmentation techniques is unlikely to be sufficient and may, in fact, aggravate bias rather than address it. To make meaningful progress, it is crucial to recognize the inherent limitations of DA and its inability to circumvent the fundamental issue of inadequate real data in certain contexts. Identifying and quantifying the data scarcity thresholds beyond which augmentation alone becomes ineffective would represent a significant step forward.

By addressing these challenges and adhering to strict train-test separation principles, we can ensure a rigorous and reliable evaluation of synthetic data generation methods in sensitive domains, ultimately enabling the development of more effective and trustworthy DA techniques for applications with limited and scarce data.

RQ2: Can the combination of XAI and HCI in generative models effectively enhance the quality and robustness of new data?
The integration of XAI into DA is still limited. Apart from XAI-guided classification tasks, some work [23] has proposed XAI-guided generation pipelines but has focused narrowly on GANs. In general, XAI aims to enhance understanding of AI systems' outputs and decision-making, improving human interpretability. Within the domain of DA, XAI may provide insights into existing methods, analyze their impact on model performance, and even guide the development of new, robust augmentation strategies with the help of domain experts bridging the gap between privacy, fairness and generation of meaningful synthetic data.

5. Conclusions

Deep generative models for medical image augmentation address the fundamental challenge of scarce training data in healthcare applications. While traditional augmentation provides some benefits, its efficacy remains limited. This research focuses on progressing from foundational models like GANs and VAEs to more advanced techniques including diffusion models and SSMs. Key advantages of generative augmentation include producing realistic synthetic data and capturing the true underlying distribution. However, limitations persist. The dual objectives of this work are: (i) employing cutting-edge advances to push beyond the current SotA in medical image synthesis, and (ii) addressing the interplay between privacy, fairness, and the generation of meaningful synthetic data by leveraging XAI and HCI for enhanced robustness.

References

[1] K. S. Button, J. P. Ioannidis, C. Mokrysz, Nosek., Power failure: why small sample size undermines the reliability of neuroscience, Nature reviews neuroscience (2013).

[2] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of big data (2019).

[3] G. Litjens, T. Kooi, B. Bejnordi, et al., A survey on deep learning in medical image analysis, Medical image analysis (2017).

[4] F. Garcea, A. Serra, F. Lamberti, L. Morra, Data augmentation for medical imaging: A systematic literature review, Computers in Biology and Medicine 152 (2023) 106391.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: NIPS, 2014.

[6] V. Sandfort, et al, Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks, Scientific reports (2019).

[7] Y. Chen, X.-H. Yang, Z. Wei, A. A. Heidari, et al., Generative adversarial networks in medical image augmentation: A review, Computers in Biology and Medicine (2022).

[8] L. Mescheder, A. Geiger, S. Nowozin, Which training methods for gans do actually converge?, in: International conference on machine learning, PMLR, 2018, pp. 3481–3490.

[9] J. P. Cohen, M. Luck, S. Honari, Distribution matching losses can hallucinate features in medical image translation, in: MICCAI, 2018.

[10] D. P. Kingma, M. Welling, Auto-encoding variational bayes, ICLR (2014).

[11] I. Goodfellow, Generative adversarial networks, Nips 2016 tutorial (2016).

[12] C. Chadebec, E. Thibeau-Sutre, N. Burgos, S. Allassonnière, Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[13] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, NIPS (2020).

[14] T. Salimans, J. Ho, Progressive distillation for fast sampling of diffusion models, arXiv preprint arXiv:2202.00512 (2022).

[15] P. Pernias, D. Rampas, M. L. Richter, C. J. Pal, M. Aubreville, Wuerstchen: An efficient architecture for large-scale text-to-image diffusion models, 2023. arXiv:2306.00637.

[16] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, 2023.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017.

[18] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision mamba: Efficient visual representation learning with bidirectional state space model, CoRR (2024).

[19] Y. Liu, Y. Tian, Y. Zhao, H. Yu, et al., Vmamba: Visual state space model, CoRR (2024).

[20] J.-H. Park, Y.-J. Ju, S.-W. Lee, Explaining generative diffusion models via visual analysis for interpretable decision-making process, Expert Systems with Applications (2024).

[21] P. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences (2017).

[22] K. Yang, J. Tao, J. Lyu, C. Ge, J. Chen, Q. Li, W. Shen, X. Zhu, X. Li, Using human feedback to fine-tune diffusion models without any reward model, CVPR (2023).

[23] S. Narteni, V. Orani, E. Ferrari, D. Verda, E. Cambiaso, M. Mongelli, A new xai-based evaluation of generative adversarial networks for imu data augmentation, 2022.