Carez Data Engine: Generation of Medical Image Datasets at Scale

Ali Rouzbayani, Rayan Sadri
Carez AI

Introduction

Developing AI for emerging imaging modalities **and** novel applications within existing modalities comes with a fundamental obstacle: **data scarcity**. Often, there are not enough patient scans available to train robust machine learning models. Furthermore, the logistical and regulatory hurdles of collecting, sharing and annotating real clinical data limit access to high quality datasets. The result is a bottleneck that slows time-to-market and hampers product validation.

Traditional data augmentation methods do not learn the underlying statistical or anatomical properties of real patient data and fail to capture genuine nuances. They tend to produce variations of the same core images and will not lead to generalized models.

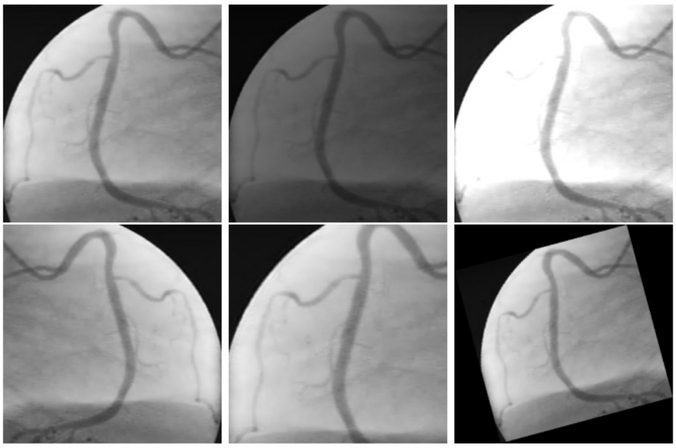

Figure 1 - Traditional data augmentation yield minimal variation, risking overfitting and poor generalization.

Solution overview

Carez AI accelerates data availability by generating large volumes of clinically valid, privacy-safe images. It removes data acquisition delays, reduces annotation work, and time-to-market.

Carez AI takes in users' datasets and uses generative modeling to create new, annotated samples at scale. Users can incorporate custom constraints, such as specific physical parameters (e.g., contrast media or radiation dosage) or anatomical structures, ensuring the generated images reflect real-world conditions.

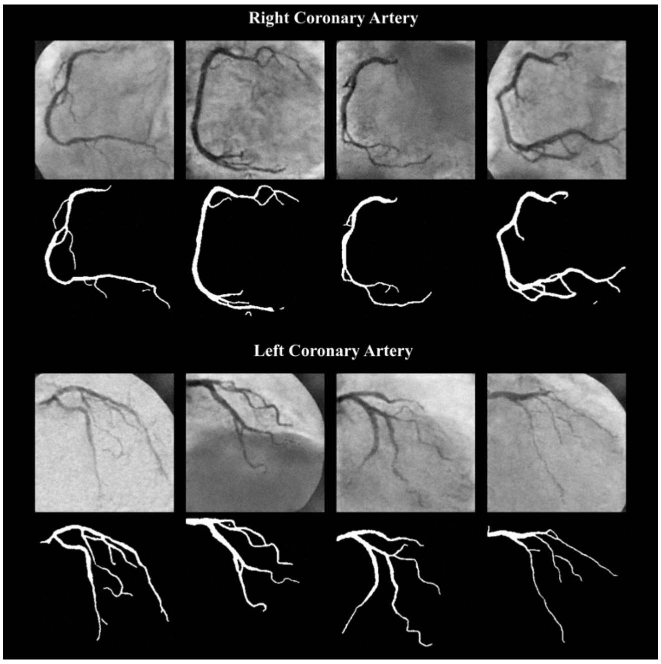

Figure 2 - **Generated using Carez AI** – synthetic coronary angiograms that follow heart anatomy.

Case study 1 : Optimization of contrast media injection to prevent permanent kidney injury

- **Goal**: Develop a vessel segmentation model that remains robust with low contrast media dosages, thereby enhancing patient safety by reducing the risk of kidney injury.

- **Data Scarcity Challenge**: Repeatedly injecting a patient with varying contrast dosages is neither safe nor feasible, leaving the dataset too limited.

- **Solution with Carez AI**: Generation of annotated angiograms at varying contrast levels

- This significantly expanded the training set and boosted segmentation accuracy as tested on original validation set without months of additional data collection.

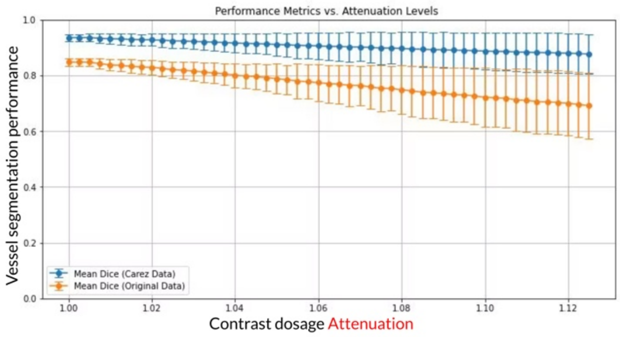

Figure 3 - Models trained with Carez AI data (blue) remain more robust than those trained on original data (orange).

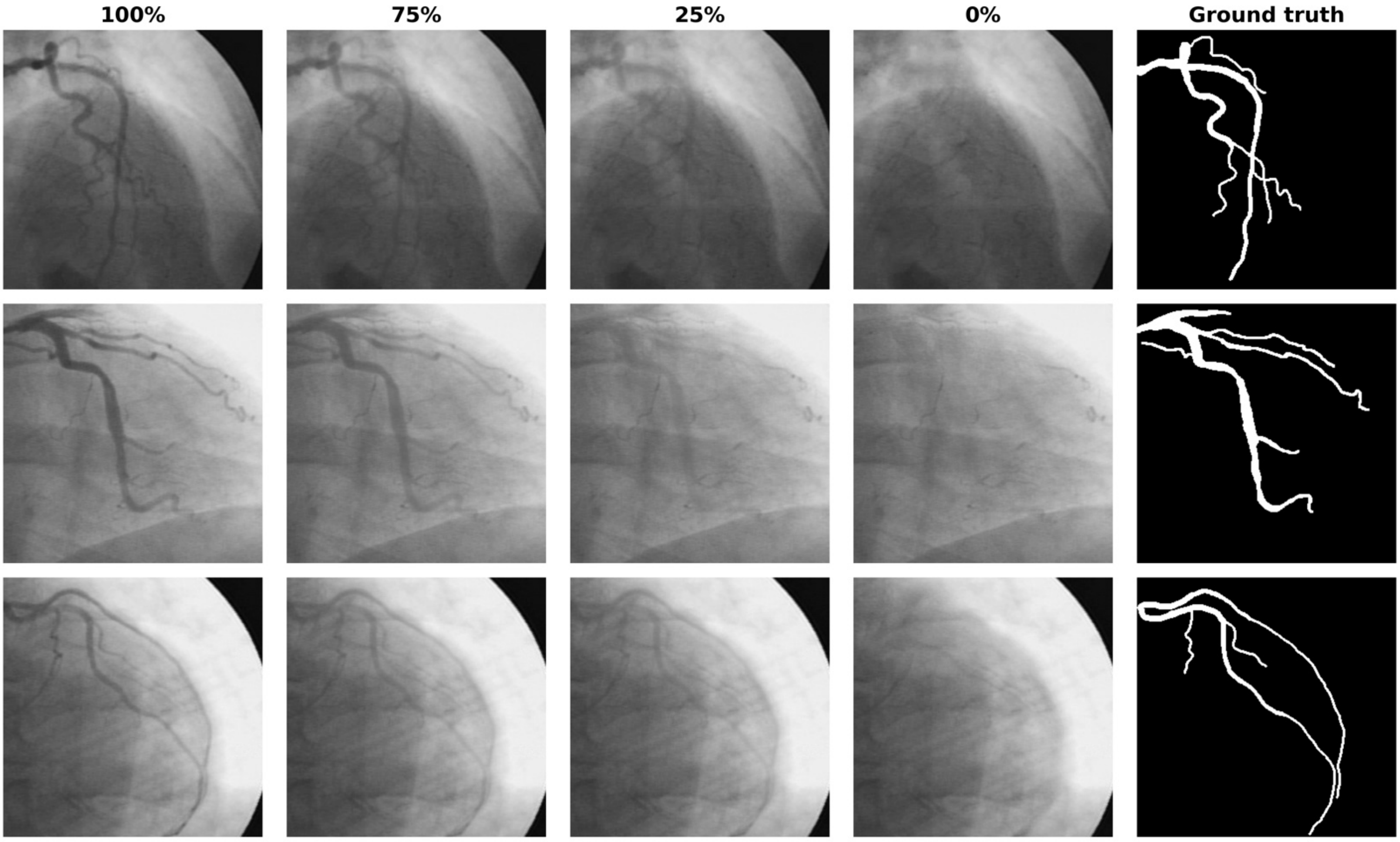

Figure 4 - **Generated using Carez AI** - Angiograms at different contrast dosage levels, illustrating anatomically consistent variations.

Case study 2: Identifying malignant vs benign breast lesions with minimal ultrasound data

- **Goal:** Enhance diagnostic accuracy for breast ultrasound scans by effectively distinguishing among benign, malignant, and normal tissue types.

- **Data Scarcity Challenge:** The dataset is heavily skewed, leaving malignant classes underrepresented. This imbalance led to poor performance for those rarer categories.

- **Solution with Carez AI:** Generating additional samples—and precisely placing synthetic tumors in underperforming image regions model’s ability to distinguish among all tissue types.

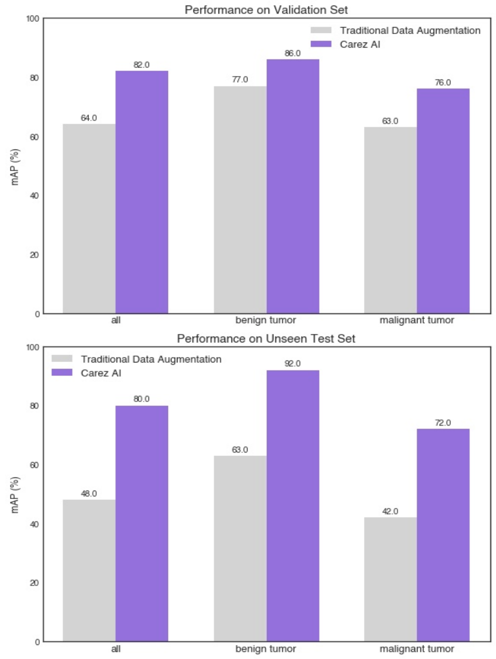

Figure 5- Bar plot comparing mAP (Mean Average Precision) for two classes. Traditional augmentation methods may show high performance on limited training data, but their metrics drop on larger, real-world test sets due to poor generalizability. In contrast, models trained with Carez AI maintain their performance on external datasets, demonstrating superior generalization.

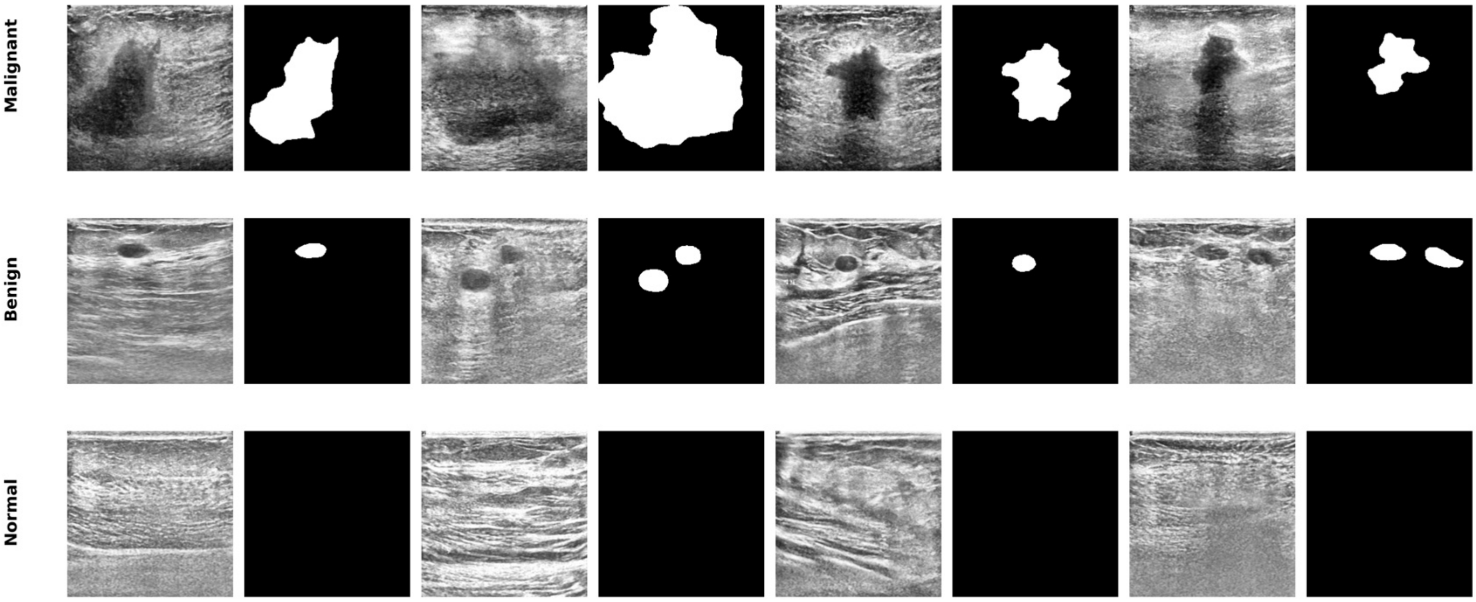

Figure 6 - **Generated using Carez AI** - Ultrasound images with annotation masks, with user-specified tumor placement.

User flow

1- Upload data and conditions

Users drag and drop the target dataset and provide details about **anatomy, physical Settings (i.e.** contrast or exposure), **label Prompts:** Include short markers (e.g., tumor location) and click Start training.

2- Select the newly created generator

After receiving a notification, the custom data generator is ready and appears on the dashboard. This process can take as little as 30 minutes for well-defined scenarios, or require a more collaborative approach for more intricate cases.

3- Data generation

With the generator selected, users specify how many images they want to produce. The platform then rapidly generates synthetic images with accurate annotations (Approximately 30 – 60 seconds per sample). The data can be exported in the desired format (e.g., COCO, binary masks, etc.).

Regulatory & ethical considerations

Privacy compliance and data anonymization: Synthetic data generation offers a massive advantage in terms of patient data regulation. Unlike traditional anonymization methods that modify existing data our approach generates data from scratch. This process retains the underlying data distribution while ensuring that no one-to-one relationship exists with any actual patient. The result is data that is both clinically relevant and inherently privacy-safe, eliminating the risk of re-identification. This high level of anonymity enables seamless data sharing and collaboration, free from the burdens typically associated with regulatory approval and data protection concerns.

Ethical use & transparency: We are committed to ensuring that synthetic data is used ethically and responsibly. Our synthetic datasets are clearly labeled and designed to complement real patient data—not to replace it. This transparency supports best practices in clinical AI pipelines and fosters trust by allowing users to easily distinguish between real and generated data. Moreover, by maintaining the integrity of the underlying clinical patterns, our data helps to build more robust and explainable AI systems, ensuring that performance improvements are grounded in realistic scenarios.

Useful resources

Khosravi, Bardia, et al. "Synthetically enhanced: unveiling synthetic data's potential in medical imaging research." *EBioMedicine* 104 (2024).

Giuffrè, Mauro, and Dennis L. Shung. "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy." *NPJ digital medicine* 6.1 (2023): 186.

Arora, Anmol. "Synthetic data: the future of open-access health-care datasets?." *The Lancet* 401.10381 (2023): 997.