Learning to generate realistic medical images to improve pancreatic cancer segmentation*

Zhuohao Tan and Scott Spurlock
[1]Department of Computer Science
Elon University
Elon, NC 27244
{ztan2, sspurlock}@elon.edu

Abstract

Pancreatic cancer is highly lethal due to the challenges associated with early-stage detection. Traditional diagnostic methods, such as imaging and biopsies, are complicated by the deep location of the pancreas within the body and the asymptomatic nature of early-stage tumors. Machine learning has emerged as a potential tool for early identification by recognizing specific patterns and features from medical images. A key first step in detecting the presence of pancreatic cancer is pancreas segmentation, or accurately delineating which pixels in a medical image correspond to a patient's pancreas. However, training accurate machine-learning segmentation models is hindered by current medical datasets, which are often limited, biased towards advanced stages of the disease, and subject to privacy restrictions, making them challenging to access and use effectively in research. This study aims to improve pancreas segmentation models (and thus ultimately cancer detection) by augmenting existing medical datasets with generated synthetic medical images that closely resemble real-world pancreatic cancer computed tomography (CT) scans. Our experiments show that adding synthetic examples to existing datasets can lead to more accurate segmentation models compared with training on real data alone.

1 Introduction

Recent advancements in machine learning, and deep neural networks (DNNs) in particular, have led to great strides in computer vision. Computer vision models are now widely applied across various domains, including autonomous driving, object detection, and image generation. Despite these successes, substantial challenges remain. Unlike traditional tabular data, images typically pose greater complexity in acquisition and often carry biases that are inherently difficult to address. Particularly with respect to medical images, such as X-rays and CT scans, available datasets are often limited due to data privacy concerns, government regulations, the cost of specialist annotations, and the relative rarity of diseases such as pancreatic cancer.

To address these limitations, this research investigates using synthetic CT scans to mitigate privacy issues, generate diverse and scalable datasets, and enhance the performance of segmentation models applied to medical imaging. Specifically, our study focuses on pancreatic cancer, a disease characterized by a high mortality rate primarily due to limited early-stage detection and insufficient data availability. Effective diagnosis of pancreatic cancer relies significantly on the accurate segmentation of the pancreas from medical images, which subsequently enables classification models to identify the presence or absence of tumors. Publicly available pancreatic cancer datasets typically have small numbers of examples, and are biased towards mid- to late-stage diagnoses, reflecting the symptomatic nature of the disease in advanced stages. Such stage-specific biases significantly hinder the training and effectiveness of AI-driven segmentation models aimed at early detection. High-quality annotated datasets necessary for effectively training pancreatic segmentation models are scarce, exacerbating the difficulty of leveraging machine learning in early-stage identification.

The objective of this research is to alleviate the challenge of data scarcity and enhance pancreas segmentation model performance by incorporating synthetic CT scans. Synthetic data will be generated using various deep neural network models, including Variational Autoencoders (VAE), Generative Adversarial Networks (GANs), and Diffusion Models (DM). We will evaluate the quality of the synthetic data both qualitatively, through visual inspection, and quantitatively, by evaluating pancreas segmentation models trained with synthetic data.

Next we will review other recent work in this area, followed by Section 3, where we describe our approach to training DNN models for data generation and for segmentation. In Section 4 we review our experiments and findings, and conclude in Section 5.

2 Related Work

Pancreas segmentation from CT images remains a challenging task due to anatomical variability, organ size, and the limited availability of annotated datasets. Prior work has addressed these challenges using deep learning-based semantic segmentation approaches. For example, convolutional neural networks (CNNs) have demonstrated success in semantic segmentation of pancreatic medical images by learning hierarchical features that capture spatial structure [6]. Building upon CNNs, AX-Unet introduced an attention-guided extension of the UNet architecture, achieving improved performance in segmenting pancreatic tumors through better spatial attention mechanisms [11].

Recent advancements in generative modeling have also contributed to early detection strategies. For instance, Li et al. [8] proposed a synthetic tumor generation approach that aims to support diagnosis by synthesizing tumor regions without requiring manual labeling. Their pipeline generates full CT volumes with simulated tumors and uses a discriminator model to predict whether a tumor is present and where it is located. While this method is innovative in simulating tumor growth for classification and localization, it does not address segmentation challenges, particularly the extraction of anatomical boundaries needed for downstream tasks like volumetric measurement or treatment planning.

In contrast, our research focuses explicitly on improving segmentation performance by generating synthetic image-label pairs for the pancreas using multiple generative models. Unlike label-free tumor synthesis [8], which generates only synthetic tumors on CT scan images, our approach ensures each synthetic image is paired with a corresponding binary mask, indicating which pixels in the image are part of a patient's pancreas, and which are not. The paired image and mask data enables supervised learning for segmentation tasks, where the goal is to predict the mask given the image. This distinction allows our synthetic data generation model to easily augment an existing pancreas segmentation data set, leading to more accurate segmentation models, which are a prerequisite for any reliable classification or detection pipeline. Moreover, we compare three types of generative models—VAE, GAN, and DM—to evaluate how different synthetic data generation techniques influence segmentation outcomes.

3 Methodology

3.1 Data

We make use of two publicly available datasets specifically designed for pancreatic CT image analysis. The Cancer Imaging Archive (TCIA) provides a

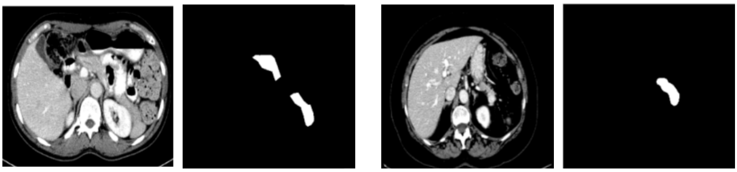

Figure 1: Example CT scan images and expert-annotated binary segmentation masks, indicating the presence of the pancreas, from the two datasets, MSD (left two) and TCIA (right two).

comprehensive source of anonymized pancreatic CT scans used widely in medical imaging research [3]. The Medical Segmentation Decathlon (MSD) is a benchmark dataset offering expert-annotated volumetric CT scans and segmentation masks for multiple organs, including the pancreas [1].

Each dataset consists of 3D volumetric CT scans stored in DICOM format, with individual files representing 2D axial slices. These slices are stacked to form full anatomical volumes. Due to the pancreas's small size and variability across patients, segmentation is especially challenging. The datasets include binary segmentation masks created by human experts, supporting supervised training despite limitations in diversity and early-stage representation. Figure 1 shows examples from the two datasets.

3.2 Data Preprocessing

We follow a standard preprocessing pipeline to prepare images for use with neural network models. The public datasets are provided in DICOM format and need to be converted into NIFTI format to facilitate 3D volumetric analysis [9]. Post-conversion, the individual 2D axial slices are grouped into uniform 4-slice sub-volumes to standardize input shapes and reduce GPU memory load during training. Any volume lacking pancreas annotation is excluded to maintain data quality. Images are intensity-normalized and resized to a fixed shape of $4 \times 256 \times 256$ (number of slices $\times$ rows $\times$ columns) to ensure consistency during model training.

3.3 Synthetic Data Generation

Using the real datasets of paired CT images and binary segmentation masks (Figure 1), we train three deep neural network (DNN) models to generate synthetic CT images given real masks. During training, real, expert-annotated binary segmentation masks are used as input to the models to condition the reconstruction process so that generated synthetic CT images correspond to

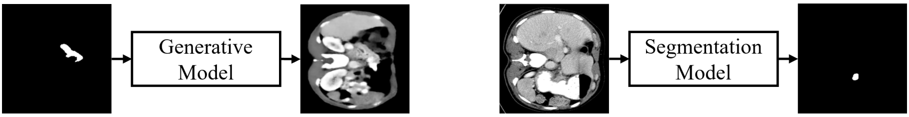

Figure 2: Each synthetic data generation model (left) is trained to take in a real binary segmentation mask and generate a corresponding synthetic CT image. A segmentation model (right) is trained to take in a CT image and output a predicted mask indicating which pixels correspond to the pancreas region.

actual masks. (See Figure 2.) Each model incorporates randomness in the generative process, so that, once a model is trained, a given real mask can be used to generate a variety of new corresponding synthetic CT images. Using each of the trained models, a synthetic dataset is generated using real binary segmentation masks as input and generating corresponding synthetic CT images. The model architectures include:

Variational Autoencoder (VAE): VAEs are probabilistic generative models that learn a compressed latent representation of input data by jointly training an encoder and decoder, using a combination of reconstruction loss and Kullback-Leibler divergence [7]. Our VAE architecture consists of encoder and decoder (both with 10 layers) with approximately 66 million trainable parameters.

Generative Adversarial Network (GAN): GANs involve a generator and discriminator trained alternately, where the generator aims to produce realistic data while the discriminator learns to distinguish real from fake samples [4]. In our implementation, the generator consists of eight transposed convolutional layers that consist of approximately 456 million trainable parameters.

Diffusion Models (DM): DMs generate samples by reversing a gradual noising process applied during training. Starting from random noise, the model iteratively denoises the input to produce realistic images [5]. Our implementation uses a 3D conditional diffusion model adapted for medical CT data generation with approximately 92 million trainable parameters. The model architecture includes a UNet-style backbone with residual blocks and skip connections, tailored for volumetric data synthesis. We configured the model with 6 residual blocks per resolution level and used a linear noise schedule across 250 diffusion timesteps.

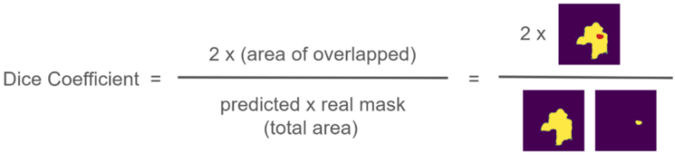

Figure 3: Example calculation for dice similarity coefficient, which measures how well two binary masks match.

3.4 Segmentation Model Training

The primary objective of the generative models described above is to produce realistic pairs of CT scan images and corresponding binary segmentation masks that can be used to train a pancreas segmentation model. As illustrated in Figure 2, the segmentation model takes a CT image as input and outputs a predicted segmentation mask that identifies the pancreas.

To perform this task, we implement a segmentation model using a 3D U-Net architecture, which has become a widely adopted baseline for volumetric medical image segmentation tasks due to its ability to capture both local and global spatial features [2]. Our implementation consists of four encoding and four decoding stages with residual skip connections at each spatial resolution level. Each stage uses 3D convolutional layers. Across all layers, the architecture contains approximately 14 million trainable parameters.

The model is trained using dice loss, a metric specifically designed to measure the overlap between predicted and ground truth segmentation masks (see Figure 3). Dice loss is particularly effective in medical imaging applications where target structures, such as the pancreas, are small and imbalanced relative to the overall image volume. By directly optimizing for spatial overlap, dice loss helps improve the model's ability to identify and accurately segment these small anatomical regions [10].

4 Results

In this section we first discuss qualitative evaluation of synthetic data produced by the three different generative models, then describe our quantitative evaluation of the synthetic data by incorporating it into training a separate segmentation model.

4.1 Qualitative Evaluation

Each of the three synthetic data generation models is trained as described in Section 3.3. Synthetic data can be evaluated qualitatively by manually com-

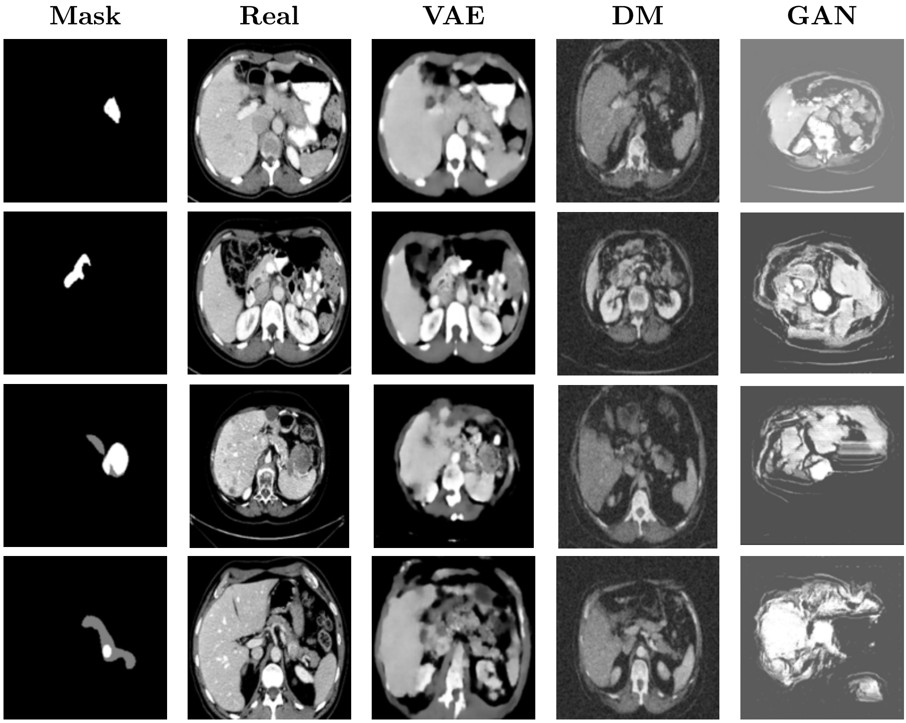

Figure 4: Example synthetic images generated by VAE, DM, and GAN models (right three columns) when given a real binary segmentation mask (left-most column) as input. For comparison, the actual (ground truth) CT scan image is shown in the second column.

paring generated images with real CT scans. Figure 4 shows sample input segmentation masks, the associated real CT image, and the synthetic CT images generated by the models. Note that the objective is to generate CT images that are realistic, but not identical to the real ones. Because our goal is ultimately to augment existing datasets, synthetic data is ideally similar to real data, but still introduces novel variation.

The images generated by the VAE model tend to show less fine detail, a tendency common to VAE image generation models. The diffusion model images appear qualitatively to be more realistic, incorporating more fine details. We also note that, for a given input mask, DM-generated images show more random variation than VAE images, suggesting that the VAE model is more likely to "memorize" individual training instances. However, DM images also include more noise, with some salt-and-pepper graininess evident throughout. The GAN model produces sharp images with the best fine details, but that overall are least realistic in terms of shape. In our experiments, the GAN model was very difficult to train successfully and had a tendency to diverge during training, as is common with GAN models, which may have contributed to the poor-quality images. Based on these results, we chose to include only the VAE and DM-based synthetic images for further evaluation.

4.2 Quantitative Evaluation

To evaluate the synthetic data quantitatively, we assess its utility for training a segmentation model, which produces a binary segmentation mask given a CT scan image. (See Figure 2.) We varied the proportion of real to synthetic data used in training the segmentation model. Due to poor qualitative results from the GAN-based samples, we limited our quantitative experiments to the following training data configurations: (1) real data only, (2) real + VAE data, (3) real + DM data, and (4) real + VAE + DM data.

The segmentation model was trained using dice loss with the Adam optimizer. Due to the volumetric nature of the data, the batch size was kept at 1 to accommodate hardware constraints. All training was conducted on a single NVIDIA RTX 4090 GPU, with 24GB of G6X memory. Training the data generation models takes approximately three weeks for the VAE and GAN models, and approximately 1.5 weeks for the DM. Training the segmentation model takes approximately 4 hours.

Segmentation model performance was assessed using the dice similarity coefficient (DSC) (see Figure 3) on a separate testing set containing exclusively real data. (Different segmentation models vary in terms of training data, but all are tested on the same real-data test set.) The baseline is a model trained on 100% real data, consisting of 3346 images. When evaluated on the test set (837 images), the baseline model achieves a DSC of **0.7606**.

Synthetic Data Source	**Synthetic : Real**		
	100% : 0%	**50% : 50%**	**25% : 75%**
VAE	0.5101	0.7664	**0.7868**
DM	0.6922	0.7641	0.7823
VAE+DM	0.6968	0.7810	0.7836
Average	0.6330	0.7705	0.7842

Table 1: Comparison of segmentation model performance in terms of dice similarity (higher is better) when trained with different proportions of synthetic and real data. The baseline performance using a model trained on purely real data is 0.7606.

To examine the impact of synthetic data on segmentation performance, we trained a series of segmentation models with varying synthetic-to-real data ratios, including 50% synthetic + 50% real, 75% synthetic + 25% real, and 100% synthetic only. Each model is evaluated on the separate test set of real examples. Table 1 shows the results.

When using 100% synthetic data (3346 images), the model trained with DM-generated images significantly outperformed the model trained with VAE-generated images (DSC of 0.6922 vs. 0.5101). Combining both VAE and DM led to a slightly improved DSC of 0.6968. Although these scores were somewhat lower than the baseline, they highlight the potential of synthetic data as a standalone resource in data-scarce environments.

At a ratio of 50% synthetic / 50% real data, both the VAE and the DM models reached DSCs comparable to the baseline (0.7664 and 0.7641, respectively). In particular, combining both synthetic types (25% VAE + 25% DM) with 50% real data achieved a higher DSC of 0.7810, surpassing the baseline of only real data. With the 25% synthetic / 75% real configuration, the VAE-only model achieved the highest overall DSC of 0.7868. The VAE+DM combination in the same setting led to a similar result. These results demonstrate that augmenting real data with synthetic data can enhance pancreas segmentation performance beyond what is achievable with real data alone. This outcome is likely due to an increase in dataset diversity because of the added synthetic data, leading to segmentation models that are less able to overfit the training data and thus more generalizable to the test set.

Figure 5 shows examples of segmentation masks produced by a segmentation model trained on a combination of real and synthetic data.

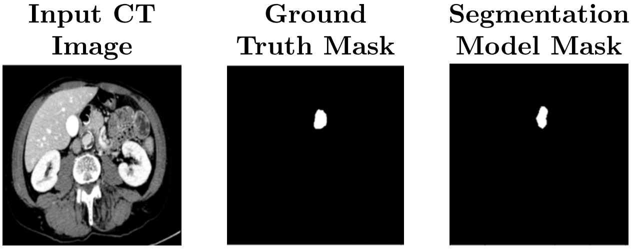

Figure 5: given the real CT image (left), corresponding to the real (ground truth) mask (center), a segmentation model trained with a mixture of real and synthetic data produces the mask on the right.

5 Conclusions and Future Work

This study demonstrates the value of incorporating synthetic medical images to improve pancreas segmentation performance in scenarios where annotated real-world data is limited. By generating synthetic CT scans (corresponding to real segmentation masks) using deep neural network generative models, we show that synthetic augmentation can significantly enhance model accuracy. Among all configurations tested, the best performing segmentation model was trained using a 25% VAE-generated synthetic and 75% real data ratio, reaching a dice similarity coefficient (DSC) of 0.7868, improving on the 0.7606 DSC baseline model trained with 100% real data. These results confirm that diverse synthetic data can improve segmentation model generalization, even surpassing the performance of models trained exclusively on real images.

An important note is that, while diffusion models generally produced more qualitatively realistic synthetic images than VAEs, neither produced perfectly realistic images. Still, when integrated with real data, both had the potential to improve segmentation model performance. This result suggests that increased data diversity helps segmentation models perform more accurately. Ultimately, our findings suggest that synthetic data may be helpful in other applications where training data is scarce, helping to mitigate the data scarcity challenge in medical imaging and potentially enabling earlier, more accurate detection of pancreatic cancer and other deadly diseases.

In the future, one promising direction may be to combine features of the different model architectures, e.g., extending the diffusion model by adding an adversarial component similar to the GAN, or incorporating a latent space similar to the VAE. We also hope to extend this work beyond segmentation to pancreatic cancer detection and further to other medical imaging tasks where more data would enable more accurate models.

References

[1] Michela Antonelli et al. “The medical segmentation decathlon”. In: *Nature communications* 13.1 (2022), p. 4128.

[2] Özgün Çiçek et al. “3D U-Net: learning dense volumetric segmentation from sparse annotation”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer. 2016, pp. 424–432.

[3] Kenneth Clark et al. “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository”. In: *Journal of digital imaging* 26 (2013), pp. 1045–1057.

[4] Ian J Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[6] Mei-Ling Huang and Yi-Zhen Wu. “Semantic segmentation of pancreatic medical images by using convolutional neural network”. In: *Biomedical Signal Processing and Control* 73 (2022), p. 103458.

[7] Durk P Kingma et al. “Semi-supervised learning with deep generative models”. In: *Advances in neural information processing systems* 27 (2014).

[8] Bowen Li et al. “Early detection and localization of pancreatic cancer by label-free tumor synthesis”. In: *arXiv preprint arXiv:2308.03008* (2023).

[9] Xiangrui Li et al. “The first step for neuroimaging data analysis: DICOM to NIfTI conversion”. In: *Journal of neuroscience methods* 264 (2016), pp. 47–56.

[10] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. “V-net: Fully convolutional neural networks for volumetric medical image segmentation”. In: *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016, pp. 565–571.

[11] Minqiang Yang et al. “AX-Unet: A deep learning framework for image segmentation to assist pancreatic tumor diagnosis”. In: *Frontiers in Oncology* 12 (2022), p. 894970.