# Data Augmentation for Medical Image Segmentation: A Comparative Analysis of Traditional Techniques and Synthetic Data Generation

**Mariana Aya S. Uchida**[1], **Erikson J. de Aguiar**[1], **Caetano Traina-Jr**[1], **Agma J. M. Traina**[1]

[1]Institute of Mathematics and Computer Science (ICMC)
*University of São Paulo (USP)* – São Carlos, Brazil

{mariaya25, erjulioaguiar}@usp.br, {caetano, agma}@icmc.usp.br

***Abstract.*** *Deep Learning has been widely applied to medical image segmentation, aiming to make structures clearer in images to help physicians identify unusual patterns and anomalies. Segmentation models face challenges in collecting a large amount of data for training, due to privacy concerns and pathological representation. Data Augmentation (DA) is an alternative to mitigate this challenge, expanding the dataset by applying transformations to the original set or creating new samples using generative methods. Despite the extensive use of DA techniques, there is still limited understanding of their relative effectiveness for medical image segmentation tasks. This work presents a method for evaluating the impact of DA methods, analyzing traditional augmentation techniques, and diffusion models for generating synthetic data in medical image segmentation models.*

## 1. Introduction

Deep learning models have provided many advances in computer vision, such as image classification, object detection, and image segmentation (Kumar et al. 2024). Medical image segmentation is crucial in medical imaging analysis because it highlights essential regions of interest, aiding physicians in disease diagnosis. Combining deep learning can make the diagnosis and treatment more accurate and efficient, such as applying the U-net model (Azad et al. 2024; Rayed et al. 2024; Ronneberger et al. 2015). Although deep learning models are successful in segmentation, they are inherently data-intensive, especially in training U-net models (Ronneberger et al. 2015). Medical image segmentation presents challenges in extracting accurate and clinically meaningful information due to small dataset size and lack of data diversity. These images often vary in intensity due to imaging equipment, settings, or patient positioning, and such inconsistencies can hinder the model learning (Rayed et al. 2024). Data augmentation has emerged as a widely adopted strategy to overcome the challenges of data availability and overfitting tendencies. This strategy aims to enhance the dataset size and diversity by introducing varied samples through transformations to the existing data (Kumar et al. 2024).

**Acknowledgments.** This study was financed in part by the São Paulo Research Foundation (FAPESP – grants 2021/08982-3, 2023/18026-8, 2024/13328-9, and 2024/18940-4), the National Research Council (CNPq), and the Coordination for Higher Education Personnel Improvement (CAPES – Finance Codes 001, 88887.914087/2023-00)

However, DA presents challenges in seeking techniques that improve application performance and reduce computational cost. Furthermore, defining the optimal sequence of application of argumentation methods and the number of augmented samples significantly impacts model performance. DA techniques can be divided into Traditional and Advanced techniques. Traditional DA comprises fundamental techniques, such as image manipulation and erasing, while advanced techniques involve innovative methods, such as mixing images and data generation. In the advanced category, diffusion DA offers a different generative technique by simulating the diffusion process, adding random noise, and smoothing to create realistic data variations. Diffusion models require high computational complexity, achieving a high performance gain across computer vision tasks, whereas traditional methods require low computational complexity, but achieve moderate performance gain (Kumar et al. 2024). Addressing the presented concerns, we proposed an approach for evaluating the impact of different DA techniques, employing traditional and diffusion-based methods in medical image segmentation. Also, we elaborated the following research questions: **RQ1:** How can the impact of traditional and diffusion-based DA methods be effective in medical image segmentation? **RQ2:** How does data quality influence synthetic data generation for segmentation model training?

## 2. Background and Related Work

The U-net model is the most widespread image segmentation architecture developed for biomedical image segmentation. It comprises two symmetrical parts: (i) the downsampling module to extract semantic and contextual features; and (ii) convolutional blocks using upsampling to gradually increase the feature maps' spatial resolutions and expand them (Azad et al. 2024; Ronneberger et al. 2015). Over the years, more than a hundred U-net methods were proposed until September 2022, and automating redundant tasks that require an expert to review large images for small features is a very clinical application of the U-net (Azad et al. 2024). Other works in the literature evaluate the use of deep learning in medical image segmentation. Joshi et al. (2024) created a framework for fundus image segmentation and retinal disease classification, achieving a Dice coefficient of 89.9%, , outperforming the original U-net architecture's 85.09% (Joshi et al. 2024). The work mentions that DA was used to prevent overfitting, however, the specific augmentation techniques were not mentioned. To assess the impact of DA in the medical image field, in an eye fundus classification task, (Goceri 2023) achieved an F1-score of 88.52% when applying a combination of color shifting, sharpening, and contrast changing to augment the inputs, and 84.19% when experimenting with rotations. Other works in the literature study data balancing techniques, such as undersampling, oversampling, and bio-inspired algorithms to augment datasets (Cavalcanti et al. 2024; Laheras et al. 2021), demonstrating the many techniques developed to tackle the data scarcity challenge presented in research fields other than the medical scenery. Generative DA methods present an alternative to mitigate privacy concerns and challenges raised in using patient data in training deep learning models. A diffusion-based model was used by (Aktas et al. 2025) to augment retinography images for ocular disease diagnosis and generate samples resembling real-world scenarios.

## 3. Methodology

**Proposed approach.** This work aims to evaluate the impact of traditional and diffusion-based DA methods in medical image segmentation. Figure 1 represents the methodology

used in this work, with letters *A* to *E* indicating its respective step subsequently.

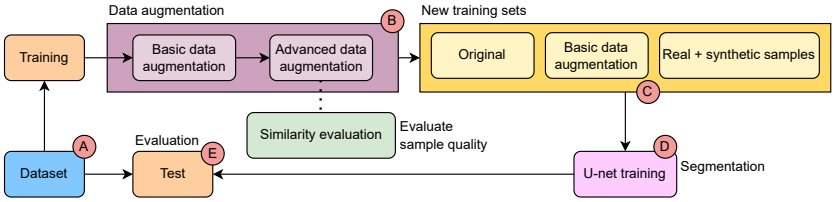


**Figure 1. The proposed methodology representation**

**Datasets.** We conducted the experiments using two datasets. The primary one is the public dataset FIVES, which consists of 800 high-resolution color fundus photographs with pixel-wise manual annotation of healthy eyes and 3 different types of eye diseases (diabetic retinopathy, glaucoma, and age-related macular degeneration), with 200 images of 2048 x 2048 pixels in each category (Jin et al. 2022). The DA experiments were conducted on the "age-related macular degeneration" label, keeping the original training and testing sets, and the training set was halved in the DA experiments to generate synthetic samples. Due to the limited sample size on the FIVES dataset, we are using the HAM10000 dataset (Codella et al. 2019; Tschandl et al. 2018) to evaluate the influence of the original dataset size and sample quality. The Melanocytic nevi (nv) label was used in the segmentation and data generation experiments, since this is the most represented label in the dataset, with 6705 images in the training set, of a size of 600 x 450 pixels (Figure 1, Part A).

**Data augmentation**. DA methods are applied to train the U-net and diffusion models in the synthetic samples generation step to mitigate the limitation of sample size (Figure 1, Part B). Traditional DA methods were applied to train diffusion and segmentation models. The original samples of both datasets were resized to 128 x 128 pixels due to computational limitations in the synthetic sample generation step. Traditional DA methods were used, since they present low computational complexity and can positively impact the model when the right combination of methods is used (Kumar et al. 2024). Therefore, rotation and color manipulation methods were used in different orders to assess the impact of order and quantity of samples necessary to train medical image segmentation models. To test the augmentation type impact, the following methods were applied to the original training split of the eye fundus dataset based on the experiments conducted by (Goceri 2023):

- *Method 1*: Seven rotation transformations were applied, preserving aspect ratio and input size. Each input went through each one of the following operations: rotation by 90º counterclockwise, rotation by 180º, rotation by 270º counterclockwise, reflection across the vertical midline, reflection across the anti-diagonal, reflection across the horizontal midline, and reflection across the main diagonal. Seven images are generated from an original image using this method.
- *Method 2*: The training set generated by Method 1 is expanded with color shifting, sharpening, and contrast changing. Color shifting was applied by randomly shifting a value for each channel with shifts in a range [-10, 10], sharpening is

used with an alpha in a range [0.1, 0.4], and contrast is used with a cutoff of zero. This method generates four images from an original input.

- *Method 3*: The training set generated by Method 1 was used to create a new training set. This method performs color-shifting, sharpening, and contrast changing. For each input, color-shifting is applied once in each RGB channel, and all three channels with shifts in a range [-10, 10]; sharpening and contrast were applied with the same parameters used in Method 2. Seven images are generated from an original input using this method.
- *Method 4*: The synthetic samples expand the original dataset. In this approach, the ground truth for synthetic samples is obtained by the U-net model, which obtained the highest segmentation performance among Methods 1 through 3. The synthetic samples were generated by training a diffusion model and applying the inpainting method to create a new image. This method involves masking a portion of the image for the model to fill in the missing part with the knowledge it acquired from the training samples. This method was used only in the segmentation experiments, and the training dataset is composed of 50% real and 50% synthetic images.

The augmentation methods were selected based on the accuracy metric achieved by (Goceri 2023), in which methods based on traditional techniques had their effects compared in the training of classification models. After the DA step, **new training sets** (Figure 1, Part C) are composed to train U-net models for **segmentation** (Figure 1, Part D), finishing the process by evaluating the models with the test set.

**Evaluation.** To evaluate the synthetic data generated by the diffusion-based DA, the Fréchet Inception Distance (FID) was used. Established by (Heusel et al. 2017), this distance captures the similarity of generated images to the real ones by measuring the distance between the feature vectors extracted from the samples. We calculated the Dice coefficient and Jaccard index to assess the segmentation performance. The Dice metric is a balanced indicator of the model's performance, the harmonic mean of accuracy and recall when used for binary segmentation maps. The Jaccard index describes the extent of overlap of the predicted segmentation mask with the ground truth. It goes between 0 and 1 and is defined as the intersection area between the predicted segmentation and the ground truth (Figure 1, Part E).

## 4. Results and Discussion

This section presents the results of image segmentation, the quality of synthetic data generated, and a discussion, including the impact of Data Augmentation (DA) on U-Net segmentation performance. The experiments were conducted on a server with NVIDIA A100 GPU using Python 3.12.2, Pytorch 2.7.0+cu126, and MONAI framework 1.4.0 (Consortium 2024). The DA techniques operated with the Albumentations library version 2.0.5 (Buslaev et al. 2020). Based on the literature, we selected the hyperparameters and trained the models, where the diffusion model was trained on 30 batches and tested on 10 batches, including a $2.5\mathrm{e}^{-5}$ learning rate. The U-net segmentation models were trained for 20 epochs, with the Adam optimizer, sigmoid function, and $1\mathrm{e}^{-4}$ learning rate.

**U-Net performance on augmentation methods**. Table 1 summarizes the metrics to evaluate segmentation performance and the impact of DA. The segmentation performance is evaluated employing the Dice and Jaccard scores, where FID evaluates the data

quality of DA. The "DA" column indicated which DA method was used to train the diffusion and segmentation model, and the arrows indicate if the metric's ideal value is closest to 0 (↓) or 1 (↑). Note that DA equal None is the baseline (without augmentation), methods 1 to 3 are traditional augmentation, and method 4 is the diffusion model.

**Table 1. Data augmentation methods evaluation by FID, mean DICE, and Jaccard**

| **Parameters** | | | **Diffusion DA** | **Segmentation** | |
|---|---|---|---|---|---|
| Dataset | DA | Samples | FID ↓ | Dice ↑ | Jaccard ↑ |
| FIVES | None | 150 | 134.22 | 0.5671 | 0.3982 |
| FIVES | Method 1 | 600 | 115.73 | 0.7741 | 0.6345 |
| FIVES | Method 2 | 2400 | 54.19 | 0.7825 | 0.6457 |
| FIVES | Method 3 | 4200 | 37.88 | 0.7886 | 0.654 |
| FIVES | Method 4 | 300 | - | 0.6668 | 0.5038 |
| HAM10000 | None | 6705 | 16.47 | 0.9430 | 0.8983 |

The segmentation experiments (See Table 1) showed that the U-net trained with the larger samples (4200) achieved the highest Dice of 0.7886. However, these results obtained by this model present improvements lower than 2% compared to the metrics achieved by the models trained with Method 2 (0.77% improvement in Dice) and Method 1 (1.87% improvement in Dice). The minimal improvement obtained by Method 3 suggests that applying color-shifting operations in each RGB channel provided minimal advancement in model learning, despite creating a training set 28 times larger than the original dataset. Method 4 shows promising results for using synthetic samples to expand the original dataset, improving the Dice coefficient by 17.58%. This leaves room for improvement by suggesting that extending the training set with more synthetic samples could benefit the model's performance. Finally, Figure 2 presents an overview of the segmentation produced by the U-Net trained on the proposed augmentation methods.

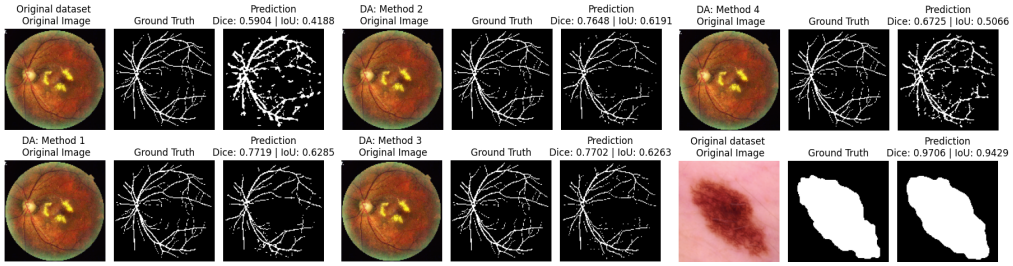


**Figure 2. Segmentation results**

**Quality analysis of synthetic data generation**. Table 1 shows that DA reduces FID between real and synthetic samples, improving the data diversity. Method 3 is the best one, achieving an FID of 37.88. The HAM10000 dataset (see Table 1) characterizes the influence of sample size in improving the FID score and consequently the data diversity. This dataset achieved an FID of 16.47, trained on the original set. To understand the impact of DA, we calculate the cosine similarity between real and synthetic samples. While the experiment demonstrates that synthetic samples closely resemble real ones, visible color differences in the inpainting region bring opportunities for further improvements. Figure 3 shows examples of synthetic images compared to real ones.

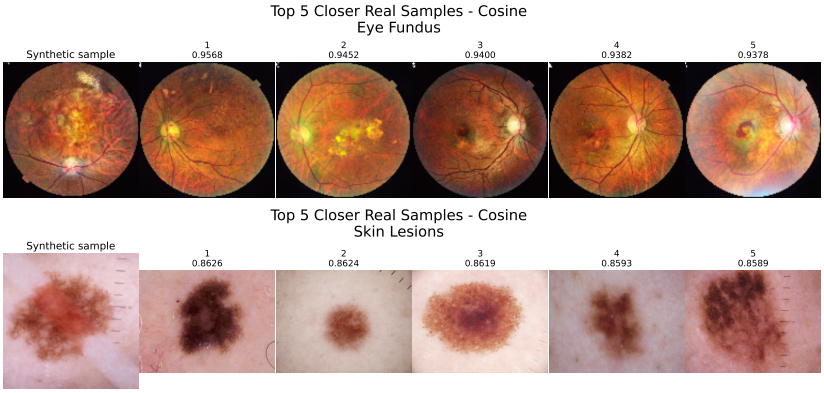


**Figure 3. Cosine similarity between synthetic and real samples**

**Discussion.** Our findings show that injecting DA in the training set can improve the segmentation performance (Dice and Jaccard) and data quality (FID). Moreover, the FID metric indicates that the data quality is enhanced according to the increase in sample size. Method 3 attained the best segmentation (highest Dice and Jaccard) and data quality scores. However, augmenting about 50% of the sample size (from method 2 to method 3) did not achieve significant outcomes in segmentation performance and requires more computational cost. One alternative to these methods is employing diffusion models to augment data.

## 5. Conclusion and Future Works

This work presented an approach to evaluate the impact of Data Augmentation (DA) methods, using diffusion methods combined with traditional ones for synthetic data generation on medical imaging segmentation. The experiments demonstrate that traditional DA and synthetic data improve the segmentation performance. However, the choice of an augmentation method and the number of generated samples directly influence the model's performance, with larger training sets providing slight but consistent improvements. Furthermore, diffusion models are promising for augmenting the dataset with realistic data, enhancing the model's data diversity and performance. As future work, we plan to extend experiments to other labels in the FIVES and HAM10000 datasets to evaluate the impact of DA on diversity and imbalanced classes. The FIVES dataset can be considered a small set, and the eye diseases represented in it present high diversity in their visual features. Also, the HAM10000 dataset presents another segmentation challenge due to its imbalanced classes and high diversity between them.

## References

[Aktas et al. 2025] Aktas, B., Ates, D. D., Duzyel, O., and Gumus, A. (2025). Diffusion-based data augmentation methodology for improved performance in ocular disease diagnosis using retinography images. *International Journal of Machine Learning and Cybernetics*, 16(5):3843–3864.

[Azad et al. 2024] Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., and Merhof, D. (2024). Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095.

[Buslaev et al. 2020] Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, 11(2).

[Cavalcanti et al. 2024] Cavalcanti, A., Brandão, D., Bezerra, E., and Coutinho, R. (2024). Avaliação de técnicas de balanceamento de dados na detecção de fraude em transações online de cartão de crédito. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 694–700, Porto Alegre, RS, Brasil. SBC.

[Codella et al. 2019] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., and Halpern, A. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic).

[Consortium 2024] Consortium, M. (2024). Monai: Medical open network for ai.

[Goceri 2023] Goceri, E. (2023). Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56(11):12561–12605.

[Heusel et al. 2017] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Jin et al. 2022] Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., and Ye, J. (2022). Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data*, 9(1):475.

[Joshi et al. 2024] Joshi, R. C., Kumar Sharma, A., and Kishore Dutta, M. (2024). Visiondeep-ai: Deep learning-based retinal blood vessels segmentation and multi-class classification framework for eye diagnosis. *Biomedical Signal Processing and Control*, 94:106273.

[Kumar et al. 2024] Kumar, T., Brennan, R., Mileo, A., and Bendechache, M. (2024). Image data augmentation approaches: A comprehensive survey and future directions. *IEEE Access*, 12:187536–187571.

[Laheras et al. 2021] Laheras, L. P., Rodrigues, P. S., Lopes, F. J. P., Palmeira, O. F. J., Falcão, A. X., Benato, B. C., and Giraldi, G. A. (2021). Aumento de dados utilizando firefly e level sets aplicado à segmentação de imagens médicas e biológicas. *Revista Eletrônica de Iniciação Científica em Computação*, 19(2).

[Rayed et al. 2024] Rayed, M. E., Islam, S. S., Niha, S. I., Jim, J. R., Kabir, M. M., and Mridha, M. (2024). Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Informatics in Medicine Unlocked*, 47:101504.

[Ronneberger et al. 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

[Tschandl et al. 2018] Tschandl, P., Rosendahl, C., and Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161.