SYNTHETIC POISONING ATTACKS: THE IMPACT OF POISONED MRI IMAGE ON U-NET BRAIN TUMOR SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning-based medical image segmentation models, such as U-Net, rely on high-quality annotated datasets to achieve accurate predictions. However, the increasing use of generative models for synthetic data augmentation introduces potential risks, particularly in the absence of rigorous quality control. In this paper, we investigate the impact of synthetic MRI data on the robustness and segmentation accuracy of U-Net models for brain tumor segmentation. Specifically, we generate synthetic T1-contrast-enhanced (T1-Ce) MRI scans using a GAN-based model with a shared encoding-decoding framework and shortest-path regularization. To quantify the effect of synthetic data contamination, we train U-Net models on progressively "poisoned" datasets, where synthetic data proportions range from 16.67% to 83.33%. Experimental results on a real MRI validation set reveal a significant performance degradation as synthetic data increases, with Dice coefficients dropping from 0.8937 (33.33% synthetic) to 0.7474 (83.33% synthetic). Accuracy and sensitivity exhibit similar downward trends, demonstrating the detrimental effect of synthetic data on segmentation robustness. These findings underscore the importance of quality control in synthetic data integration and highlight the risks of unregulated synthetic augmentation in medical image analysis. Our study provides critical insights for the development of more reliable and trustworthy AI-driven medical imaging systems.

1 INTRODUCTION

Deep learning-based segmentation models (Minaee et al., 2021), such as U-Net (Ronneberger et al., 2015), have demonstrated remarkable success in medical image analysis (Azad et al., 2024; Du et al., 2020), particularly in brain tumor segmentation tasks (Abidin et al., 2024; Ranjbarzadeh et al., 2023; Das et al., 2022; Magadza & Viriri, 2021). These models rely heavily on high-quality image-segmentation pairs to ensure accurate and reliable predictions. However, the growing adoption of generative models for synthetic medical image creation introduces new challenges (Dayarathna et al., 2024). While synthetic data can potentially augment training datasets, improve data diversity, and address class imbalances, its uncontrolled incorporation may lead to significant performance degradation (Hao et al., 2024). Without rigorous quality control, synthetic data can act as a form of "data poisoning", negatively impacting model robustness and segmentation accuracy (Yerlikaya & Bahtiyar, 2022; Pitropakis et al., 2019).

In recent years, generative adversarial networks (GANs) (Goodfellow et al., 2014; 2020) have emerged as a popular technique for generating synthetic medical images (AlAmir & AlGhamdi, 2022; Singh & Raza, 2021; Nie et al., 2017). These models leverage learned distributions from real data to synthesize realistic samples. While some studies have explored the benefits of GAN-generated data for augmentation (Makhlouf et al., 2023; Zhang et al., 2023; Chen et al., 2022a; Hatamizadeh et al., 2021), few have systematically examined the risks associated with using synthetic medical images in segmentation tasks. Specifically, the effects of synthetic data contamination on segmentation models remain insufficiently studied, raising concerns about potential accuracy degradation and unreliable clinical applications (Singkorapoom & Phoomvuthisarn, 2023).

To address such a gap, we evaluate the impact of synthetic MRI data on the performance of U-Net (Ronneberger et al., 2015) models for brain tumor segmentation. We consider synthetic data as a type of contamination and investigate how increasing proportions of synthetic data influence segmentation robustness. Using a novel GAN-based model (Xie et al., 2023), we generate synthetic T1-contrast-enhanced (T1-Ce) MRI images from paired CT-MRI datasets and introduce them into training sets at varying proportions. We compare a baseline U-Net trained exclusively on real MRI data against U-Net models trained on progressively "poisoned" datasets containing increasing amounts of synthetic data. Segmentation performance is assessed using Dice coefficient, Jaccard index, accuracy, and sensitivity to quantify the extent of degradation. The results demonstrate a significant decline in segmentation performance as the proportion of synthetic data increases, with notable drops in Dice coefficients, Jaccard index, and sensitivity. These findings emphasize the importance of establishing best practices for incorporating synthetic data in medical image segmentation pipelines. By highlighting the potential risks of synthetic data contamination, this study provides valuable insights for developing robust and trustworthy deep learning (Wang et al., 2023a; Li et al., 2023; Huang et al., 2018; Hanif et al., 2018; Li et al., 2024b; Zheng et al., 2024) in medical imaging applications (Shukla et al., 2023; Teng et al., 2024; Fidon, 2023; Shi et al., 2024).

2 Related Works

Brain Tumor Segmentation Brain tumor segmentation has been a critical task in medical image analysis, enabling precise delineation of tumor regions for diagnosis and treatment planning (Wadhwa et al., 2019; Gordillo et al., 2013). Traditional methods relied on handcrafted features (Mecheter et al., 2022; Khan et al., 2020; Hasan et al., 2019) and classical machine learning models (Soomro et al., 2022; Amin et al., 2019; Bakas et al., 2018), but deep learning approaches, particularly convolutional neural networks (CNNs) (Li et al., 2021), have significantly advanced segmentation performance (Havaei et al., 2017; Pereira et al., 2016). U-Net (Ronneberger et al., 2015) and its variants (Azad et al., 2024; Siddique et al., 2021) have become the backbone of many segmentation pipelines due to their encoder-decoder architecture and skip connections, which preserve spatial information. More recent methods leverage transformer-based architectures (Ghazouani et al., 2024; Wang et al., 2023b; Jiang et al., 2022; Huang et al., 2022) and hybrid CNN-Transformer models (Liu et al., 2024; Kang et al., 2024; Chen et al., 2022b; Jia & Shu, 2021) to enhance feature representation and improve segmentation accuracy. Despite these advancements, the robustness of segmentation models remains a concern, especially when trained on datasets with varying levels of synthetic content.

GAN-based MRI Synthesis The generation of synthetic MRI images has gained significant attention due to its potential to augment datasets, address data scarcity, and enable cross-modality learning (Tiwari et al., 2025; Choi et al., 2025; Pani & Chawla, 2024; Koetzier et al., 2024; Hamghalam & Simpson, 2024; Ji et al., 2022; Han et al., 2018; Blystad et al., 2012). Generative adversarial networks (GANs) (Goodfellow et al., 2014; 2020) and variational autoencoders (VAEs) (Kingma, 2013; Pinheiro Cinelli et al., 2021) have been widely explored for MRI synthesis (Tavse et al., 2022; Laptev et al., 2021). Conditional GANs (Mirza, 2014) and cycle-consistent GANs (Zhu et al., 2017) have been applied for modality translation, such as generating MRI from CT scans. More recent works incorporate structural constraints and perceptual losses to improve anatomical consistency in synthetic images. However, concerns persist regarding the quality and fidelity of synthetic images, as even minor artifacts or inconsistencies can propagate through downstream tasks, adversely affecting segmentation performance. In this context, synthetic medical images may act as a form of data poisoning, compromising model reliability and clinical applicability (Singkorapoom & Phoomvuthisarn, 2023).

Data Poisoning Attack Data poisoning attacks involve injecting manipulated, low-quality, or malicious data into training datasets to degrade model performance or induce adversarial vulnerabilities (Yerlikaya & Bahtiyar, 2022; Fan et al., 2022). In the medical imaging domain, poisoning can occur through mislabeled (Tolpegin et al., 2020; Lin et al., 2021), perturbed (Martinelli et al., 2023; Bortsova et al., 2021). Prior research has demonstrated that even small perturbations in training data can lead to significant performance degradation in classification and segmentation models (Szegedy, 2013; Chakraborty et al., 2021). While poisoning attacks and corresponding mitigation strategies (Goldblum et al., 2022; Schwarzschild et al., 2021; Fu et al., 2024; Li et al., 2024a) have been extensively studied in general computer vision tasks (Wei et al., 2024; Akhtar & Mian, 2018), their impact on medical imaging pipelines remains underexplored. Our work investigates the effect of synthetic MRI data as a form of data poisoning, evaluating its impact on brain tumor segmentation performance.

3 METHODS

In this study, we investigate the potential degradation in segmentation accuracy when synthetic data is incorporated into the training process, an overall workflow is shown in Figure 1. Formally, let M denote a medical AI model trained for segmentation tasks, and let S represent an image synthesis model designed to generate synthetic medical images. We define a real medical dataset as D, and a modified dataset containing synthetic samples generated using S as D'. Our objective is to analyze the effects of training M on D'.

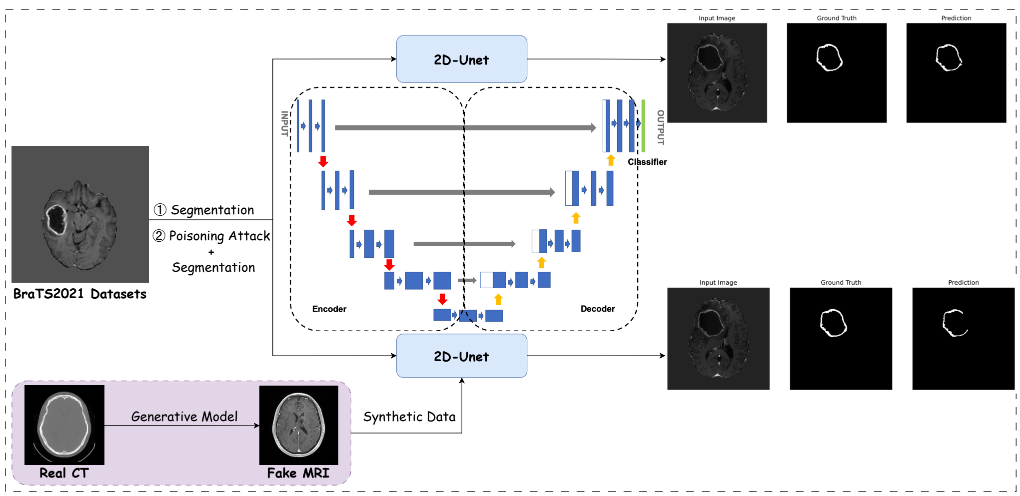

Figure 1: Overall workflow.

Specifically, as shown in Algorithm 1, our study follows these steps: (i) We first prepare a medical dataset D, and employ a generative model S (Xie et al., 2023), as shown in Figure 2, to produce synthetic data, resulting in a modified dataset D'. (ii) We then train a baseline model M trained solely on D, and a model M' trained on D', which contains synthetic images. (iii) We evaluate the segmentation performance of both models M and M' using metrics including the Dice coefficient, Jaccard index, accuracy, and sensitivity.

Algorithm 1 Training and Evaluation of U-Net with Synthetic Data

Require: $\mathcal{D} = \{(x, y)\}, S, \mathcal{P} = \{16.67\%, 33.33\%, 50.00\%, 66.67\%, 83.33\%\}$
for $p \in \mathcal{P}$ **do** ▷ Poisoning
    $\mathcal{X}' \leftarrow S(\mathcal{D})$
    $\mathcal{D}'(p) \leftarrow \mathcal{D} \cup \mathcal{X}'$
end for
$\mathcal{M} \leftarrow$ Train U-Net on $\mathcal{D}$
for $p \in \mathcal{P}$ **do** ▷ Training
    $\mathcal{M}'(p) \leftarrow$ Train U-Net on $\mathcal{D}'(p)$
end for
for $p \in \mathcal{P}$ **do** ▷ Evaluation
    Compute Dice$(\mathcal{M})$, Dice$(\mathcal{M}'(p))$
    Compute Jaccard, Accuracy, Sensitivity
end for

We expect that the segmentation performance of M' will be lower than that of M, formally expressed as: Dice$(M') <$ Dice(M), where Dice(M) and Dice(M') represent the Dice coefficients of the models trained on D and D', respectively. Through systematic experimentation, we aim to

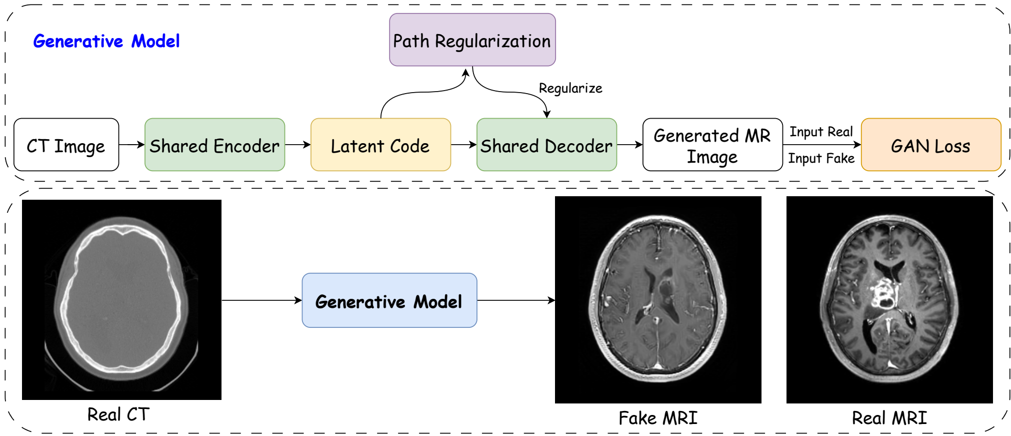

Figure 2: Workflow of generative model (Xie et al., 2023).

quantify the extent of performance degradation and provide insights into the risks associated with synthetic data contamination in medical image segmentation tasks.

4 Experiments

4.1 Setup

Dataset We conduct our experiments using the publicly available BraTS2021 dataset, which contains T1-contrast-enhanced (T1-Ce) MRI scans from 150 glioma patients along with their corresponding enhanced tumor (ET) segmentation masks (Menze et al., 2014). To introduce synthetic data, we employ a generative adversarial network (GAN) (Goodfellow et al., 2020; 2014) designed for cross-domain medical image translation (Xie et al., 2023). Specifically, the GAN is trained on 660 paired CT-MRI datasets and features a shared encoding-decoding framework with shortest-path regularization (Xie et al., 2023) to ensure anatomical consistency during translation. Some cases of synthetic MRI can be found in the appendix A. The trained model generates 150 synthetic T1-Ce MRI scans, which are then incorporated into our training set at varying proportions.

Protocol We evaluate the impact of synthetic MRI data on U-Net segmentation performance. Let $\mathcal{D} = \{(x, y)\}$ represent the original dataset, where x denotes real MRI scans and y the corresponding segmentation masks. Using the trained GAN model, we generate synthetic MRI images $\mathcal{X}'$. We construct modified datasets $\mathcal{D}'(p)$ by mixing real MRI scans with synthetic samples, where $p \in \{16.67\%, 33.33\%, 50.00\%, 66.67\%, 83.33\%\}$ represents the proportion of synthetic data. Two types of U-Net models are trained: (i) A baseline model $\mathcal{M}$ trained solely on $\mathcal{D}$, establishing a performance reference. (ii) Poisoned models $\mathcal{M}'(p)$ trained on $\mathcal{D}'(p)$, simulating different levels of synthetic data contamination.

Metrics We evaluate the segmentation performance of $\mathcal{M}$ and $\mathcal{M}'(p)$ on a real MRI test set using the following standard metrics:

(i) *Dice Coefficient*: Measures the spatial overlap between the predicted segmentation $\hat{Y}$ and the ground truth Y. Defined as:

$$\text{Dice} = \frac{2|\hat{Y} \cap Y|}{|\hat{Y}| + |Y|} \tag{1}$$

where $|\hat{Y} \cap Y|$ represents the number of correctly segmented pixels, and $|\hat{Y}|$ and $|Y|$ denote the total number of pixels in the predicted and ground truth masks, respectively. A higher Dice score indicates better segmentation performance.

(ii) *Jaccard Index*: Also known as the Intersection-over-Union (IoU), this metric evaluates the proportion of correctly segmented pixels relative to the union of predicted and ground truth segmenta-

Table 1: Quantitative description of metrics on varying poisoning rates

Poisoning Rate(%)	**Dice**	**Jaccard**	**Accuracy**	**Sensitivity**
0.00	0.8939 ± 0.1243	0.8184 ± 0.1546	0.9983 ± 0.0011	0.9136 ± 0.1578
16.67	0.8650 ± 0.2072	0.7937 ± 0.2151	0.9638 ± 0.1854	0.8790 ± 0.2390
33.33	0.8937 ± 0.0722	0.8145 ± 0.1071	0.9981 ± 0.0013	0.9191 ± 0.1173
50.00	0.8572 ± 0.1580	0.7738 ± 0.1810	0.9979 ± 0.0011	0.9292 ± 0.1208
66.67	0.8146 ± 0.2457	0.7360 ± 0.2458	0.9978 ± 0.0013	0.8328 ± 0.2817
83.33	0.7474 ± 0.2650	0.6486 ± 0.2579	0.9967 ± 0.0020	0.7577 ± 0.3054

tions:

$$\text{Jaccard} = \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|} \tag{2}$$

Jaccard provides a stricter evaluation compared to Dice, as it penalizes false positives and false negatives more severely.

(iii) *Accuracy*: Measures the overall correctness of pixel classification, considering both the segmented tumor region and the background:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Accuracy alone can be misleading in imbalanced segmentation tasks, where background pixels dominate.

(iv) *Sensitivity*: Also known as recall or true positive rate, sensitivity quantifies the model's ability to correctly identify tumor regions:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

A higher sensitivity indicates fewer missed tumor regions, which is critical for medical applications where under-segmentation could lead to misdiagnoses.

4.2 Results

Figure 3 illustrates the segmentation outputs of the U-Net model under varying poisoning rates. At low synthetic data proportions (e.g., 16.67%), model predictions remain close to the ground truth. However, as p increases, segmentation quality deteriorates, with higher poisoning levels leading to incorrect tumor boundary delineations. Table 1 presents the quantitative impact of synthetic data contamination. The Dice coefficients decrease from 0.8937 ($p = 33.33\%$) to 0.7474 ($p = 83.33\%$), confirming a strong correlation between synthetic data proportion and segmentation degradation. Jaccard and sensitivity exhibit similar trends, with significant performance drops beyond $p = 50\%$. However, the increase of a portion of the synthetic data has a negligible effect on accuracy. These findings suggest that, while low proportions of synthetic data may not drastically harm model performance, excessive reliance on synthetic data compromises segmentation robustness.

To further understand the effects of synthetic MRI data, we analyze segmentation performance across different poisoning thresholds. Our results indicate that models trained with $p \leq 33.33\%$ maintain relatively stable performance, while those with $p \geq 50\%$ suffer from severe degradation. This highlights the importance of synthetic data curation, suggesting that controlled synthetic augmentation may be feasible if appropriately regulated.

Our findings raise critical concerns about the integration of synthetic medical images in training pipelines. While synthetic MRI augmentation can be beneficial in low proportions, excessive synthetic data exposure introduces model biases and reduces segmentation reliability. These results emphasize the need for quality control mechanisms and hybrid training strategies that combine real and synthetic data optimally to mitigate potential risks in medical AI applications.

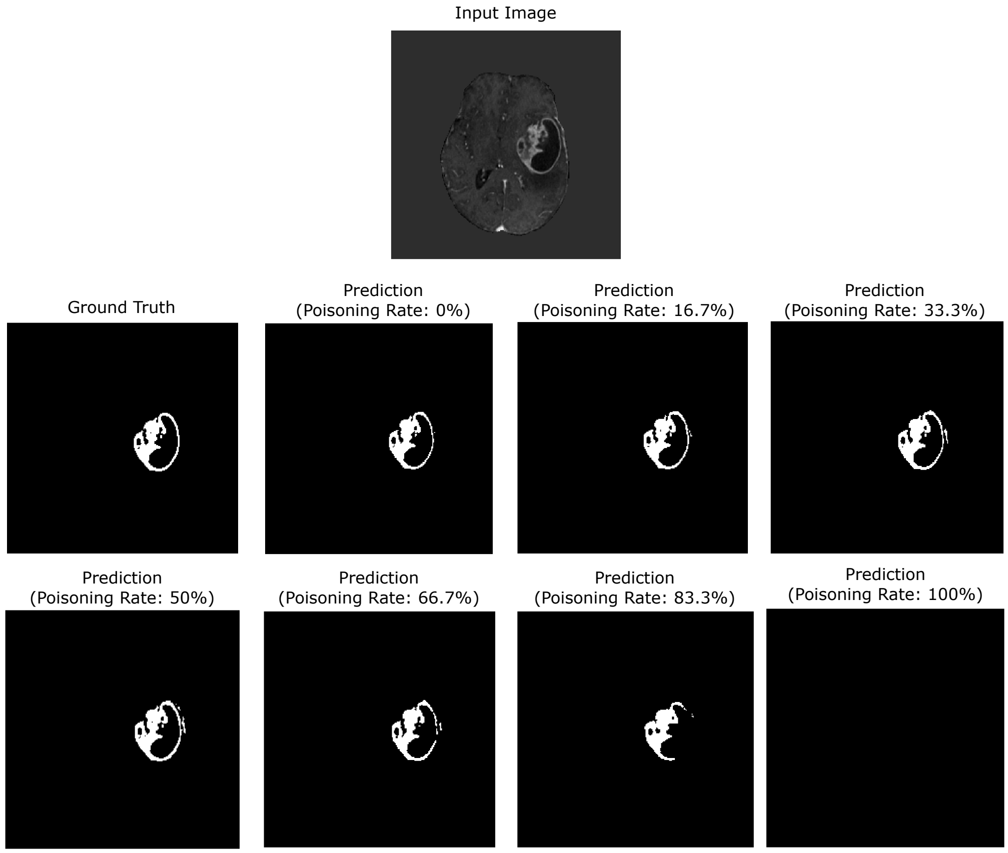

Figure 3: A sample of segmentation results of the ET region from the same MRI scan using the U-Net model at varying poisoning rates.

5 Conclusion

We investigated the impact of synthetic MRI data on the robustness and segmentation accuracy of U-Net models for brain tumor segmentation. Experiment results suggest that the inclusion of synthetic data, when not properly regulated, significantly degrades segmentation performance. As the proportion of synthetic MRI data increased, we observed a substantial decline in key evaluation metrics, including Dice coefficient, Jaccard index, accuracy, and sensitivity. Our findings highlight that while small proportions of synthetic data may not drastically impair model performance, excessive reliance on synthetic samples introduces severe biases, compromises segmentation reliability, and leads to inaccurate tumor boundary delineations. We provide crucial insights for designing safer, more reliable deep learning models in medical imaging. As the adoption of generative models continues to expand, our work serves as a foundation for establishing best practices in the responsible integration of synthetic data in AI-driven healthcare systems.

References

Zain Ul Abidin, Rizwan Ali Naqvi, Amir Haider, Hyung Seok Kim, Daesik Jeong, and Seung Won Lee. Recent deep learning-based brain tumor segmentation models using multi-modality magnetic resonance imaging: a prospective survey. *Frontiers in Bioengineering and Biotechnology*, 12: 1392807, 2024.

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

Manal AlAmir and Manal AlGhamdi. The role of generative adversarial network in medical image analysis: An in-depth survey. *ACM Computing Surveys*, 55(5):1–36, 2022.

Javaria Amin, Muhammad Sharif, Mudassar Raza, Tanzila Saba, and Muhammad Almas Anjum. Brain tumor detection using statistical and machine learning method. *Computer methods and programs in biomedicine*, 177:69–79, 2019.

Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Ida Blystad, Jan Bertus Marcel Warntjes, O Smedby, Anne-Marie Landtblom, Peter Lundberg, and Elna-Marie Larsson. Synthetic mri of the brain in a clinical setting. *Acta radiologica*, 53(10): 1158–1163, 2012.

Gerda Bortsova, Cristina González-Gonzalo, Suzanne C Wetstein, Florian Dubost, Ioannis Katramados, Laurens Hogeweg, Bart Liefers, Bram van Ginneken, Josien PW Pluim, Mitko Veta, et al. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*, 73:102141, 2021.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021.

Yizhou Chen, Xu-Hua Yang, Zihan Wei, Ali Asghar Heidari, Nenggan Zheng, Zhicheng Li, Huiling Chen, Haigen Hu, Qianwei Zhou, and Qiu Guan. Generative adversarial networks in medical image augmentation: a review. *Computers in Biology and Medicine*, 144:105382, 2022a.

Yu Chen, Ming Yin, Yu Li, and Qian Cai. Csu-net: A cnn-transformer parallel network for multimodal brain tumour segmentation. *Electronics*, 11(14):2226, 2022b.

Yangsean Choi, Ji Su Ko, Ji Eun Park, Geunu Jeong, Minkook Seo, Yohan Jun, Shohei Fujita, and Berkin Bilgic. Beyond the conventional structural mri: Clinical application of deep learning image reconstruction and synthetic mri of the brain. *Investigative Radiology*, 60(1):27–42, 2025.

Suchismita Das, Gopal Krishna Nayak, Luca Saba, Mannudeep Kalra, Jasjit S Suri, and Sanjay Saxena. An artificial intelligence framework and its bias for brain tumor segmentation: A narrative review. *Computers in biology and medicine*, 143:105273, 2022.

Sanuwani Dayarathna, Kh Tohidul Islam, Sergio Uribe, Guang Yang, Munawar Hayat, and Zhaolin Chen. Deep learning based synthesis of mri, ct and pet: Review and analysis. *Medical image analysis*, 92:103046, 2024.

Getao Du, Xu Cao, Jimin Liang, Xueli Chen, and Yonghua Zhan. Medical image segmentation based on u-net: A review. *Journal of Imaging Science & Technology*, 64(2), 2020.

Jiaxin Fan, Qi Yan, Mohan Li, Guanqun Qu, and Yang Xiao. A survey on data poisoning attacks and defenses. In *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)*, pp. 48–55. IEEE, 2022.

Lucas Fidon. Trustworthy deep learning for medical image segmentation. *arXiv preprint arXiv:2305.17456*, 2023.

Tingchen Fu, Mrinank Sharma, Philip Torr, Shay B Cohen, David Krueger, and Fazl Barez. Poisonbench: Assessing large language model vulnerability to data poisoning. *arXiv preprint arXiv:2410.08811*, 2024.

Fethi Ghazouani, Pierre Vera, and Su Ruan. Efficient brain tumor segmentation using swin transformer and enhanced local self-attention. *International Journal of Computer Assisted Radiology and Surgery*, 19(2):273–281, 2024.

Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Nelly Gordillo, Eduard Montseny, and Pilar Sobrevilla. State of the art survey on mri brain tumor segmentation. *Magnetic resonance imaging*, 31(8):1426–1438, 2013.

Mohammad Hamghalam and Amber L Simpson. Medical image synthesis via conditional gans: Application to segmenting brain tumours. *Computers in Biology and Medicine*, 170:107982, 2024.

Changhee Han, Hideaki Hayashi, Leonardo Rundo, Ryosuke Araki, Wataru Shimoda, Shinichi Muramatsu, Yujiro Furukawa, Giancarlo Mauri, and Hideki Nakayama. Gan-based synthetic brain mr image generation. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 734–738. IEEE, 2018.

Muhammad Abdullah Hanif, Faiq Khalid, Rachmad Vidya Wicaksana Putra, Semeen Rehman, and Muhammad Shafique. Robust machine learning systems: Reliability and security for deep neural networks. In *2018 IEEE 24th international symposium on on-line testing and robust system design (IOLTS)*, pp. 257–260. IEEE, 2018.

Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. Synthetic data in ai: Challenges, applications, and ethical implications. *arXiv preprint arXiv:2401.01629*, 2024.

Ali M Hasan, Hamid A Jalab, Farid Meziane, Hasan Kahtan, and Ahmad Salah Al-Ahmad. Combining deep and handcrafted image features for mri brain scan classification. *IEEE Access*, 7: 79959–79967, 2019.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284. Springer, 2021.

Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

Liqun Huang, Enjun Zhu, Long Chen, Zhaoyang Wang, Senchun Chai, and Baihai Zhang. A transformer-based generative adversarial network for brain tumor segmentation. *Frontiers in Neuroscience*, 16:1054948, 2022.

Xiaowei Huang, Daniel Kroening, Marta Kwiatkowska, Wenjie Ruan, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. Safety and trustworthiness of deep neural networks: A survey. *arXiv preprint arXiv:1812.08342*, pp. 151, 2018.

Sooyeon Ji, Dongjin Yang, Jongho Lee, Seung Hong Choi, Hyeonjin Kim, and Koung Mi Kang. Synthetic mri: technologies and applications in neuroradiology. *Journal of Magnetic Resonance Imaging*, 55(4):1013–1025, 2022.

Qiran Jia and Hai Shu. Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pp. 3–14. Springer, 2021.

Yun Jiang, Yuan Zhang, Xin Lin, Jinkun Dong, Tongtong Cheng, and Jing Liang. Swinbts: A method for 3d multimodal brain tumor segmentation using swin transformer. *Brain sciences*, 12 (6):797, 2022.

Ming Kang, Fung Fung Ting, Raphaël C-W Phan, Zongyuan Ge, and Chee-Ming Ting. A multimodal feature distillation with cnn-transformer network for brain tumor segmentation with incomplete modalities. *arXiv preprint arXiv:2404.14019*, 2024.

Hikmat Khan, Pir Masoom Shah, Munam Ali Shah, Saif ul Islam, and Joel JPC Rodrigues. Cascading handcrafted features and convolutional neural network for iot-enabled brain tumor segmentation. *Computer communications*, 153:196–207, 2020.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lennart R Koetzier, Jie Wu, Domenico Mastrodicasa, Aline Lutz, Matthew Chung, W Adam Koszek, Jayanth Pratap, Akshay S Chaudhari, Pranav Rajpurkar, Matthew P Lungren, et al. Generating synthetic data for medical imaging. *Radiology*, 312(3):e232471, 2024.

Vladislav V Laptev, Olga M Gerget, and Nataliia A Markova. Generative models based on vae and gan for new medical data synthesis. *Society 5.0: Cyberspace for advanced human-centered society*, pp. 217–226, 2021.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

Tianhao Li, Jingyu Lu, Chuangxin Chu, Tianyu Zeng, Yujia Zheng, Mei Li, Haotian Huang, Bin Wu, Zuoxian Liu, Kai Ma, et al. Scisafeeval: a comprehensive benchmark for safety alignment of large language models in scientific tasks. *arXiv preprint arXiv:2410.03769*, 2024a.

Tianhao Li, Yujia Zheng, Weizhi Ma, Guangshuo Wang, Zhengping Li, and Lijun Wang. P-4.33: Trustworthy metaverse: A comprehensive investigation into security risks and privacy issues in artificial intelligence-extended reality systems. In *SID Symposium Digest of Technical Papers*, volume 55, pp. 872–877. Wiley Online Library, 2024b.

Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.

Jing Lin, Ryan Luley, and Kaiqi Xiong. Active learning under malicious mislabeling and poisoning attacks. In *2021 IEEE global communications conference (GLOBECOM)*, pp. 1–6. IEEE, 2021.

Yu Liu, Yize Ma, Zhiqin Zhu, Juan Cheng, and Xun Chen. Transsea: Hybrid cnn-transformer with semantic awareness for 3d brain tumor segmentation. *IEEE Transactions on Instrumentation and Measurement*, 2024.

Tirivangani Magadza and Serestina Viriri. Deep learning for brain tumor segmentation: a survey of state-of-the-art. *Journal of Imaging*, 7(2):19, 2021.

Ahmed Makhlouf, Marina Maayah, Nada Abughanam, and Cagatay Catal. The use of generative adversarial networks in medical image augmentation. *Neural Computing and Applications*, 35 (34):24055–24068, 2023.

Fabio Martinelli, Francesco Mercaldo, Marcello Di Giammarco, and Antonella Santone. Data poisoning attacks over diabetic retinopathy images classification. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 3698–3703. IEEE, 2023.

Imene Mecheter, Maysam Abbod, Abbes Amira, and Habib Zaidi. Deep learning with multiresolution handcrafted features for brain mri segmentation. *Artificial intelligence in medicine*, 131: 102365, 2022.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2014.

Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

Mehdi Mirza. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pp. 417–425. Springer, 2017.

Kaliprasad Pani and Indu Chawla. Synthetic mri in action: A novel framework in data augmentation strategies for robust multi-modal brain tumor segmentation. *Computers in Biology and Medicine*, 183:109273, 2024.

Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5): 1240–1251, 2016.

Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Antúnio Barros da Silva, and Sérgio Lima Netto. Variational autoencoder. In *Variational Methods for Machine Learning with Applications to Deep Networks*, pp. 111–149. Springer, 2021.

Nikolaos Pitropakis, Emmanouil Panaousis, Thanassis Giannetsos, Eleftherios Anastasiadis, and George Loukas. A taxonomy and survey of attacks against machine learning. *Computer Science Review*, 34:100199, 2019.

Ramin Ranjbarzadeh, Annalina Caputo, Erfan Babaee Tirkolaee, Saeid Jafarzadeh Ghoushchi, and Malika Bendechache. Brain tumor segmentation of mri images: A comprehensive review on the application of artificial intelligence tools. *Computers in biology and medicine*, 152:106405, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pp. 9389–9398. PMLR, 2021.

Congzhen Shi, Ryan Rezai, Jiaxi Yang, Qi Dou, and Xiaoxiao Li. A survey on trustworthiness in foundation models for medical image analysis. *arXiv preprint arXiv:2407.15851*, 2024.

Sneha Shukla, Lokendra Birla, Anup Kumar Gupta, and Puneet Gupta. Trustworthy medical image segmentation with improved performance for in-distribution samples. *Neural Networks*, 166: 127–136, 2023.

Nahian Siddique, Sidike Paheding, Colin P Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE access*, 9:82031–82057, 2021.

Nripendra Kumar Singh and Khalid Raza. Medical image generation using generative adversarial networks: A review. *Health informatics: A computational perspective in healthcare*, pp. 77–96, 2021.

Pakpoom Singkorapoom and Suronapee Phoomvuthisarn. Pre-trained model robustness against gan-based poisoning attack in medical imaging analysis. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 302–313. Springer, 2023.

Toufique A Soomro, Lihong Zheng, Ahmed J Afifi, Ahmed Ali, Shafiullah Soomro, Ming Yin, and Junbin Gao. Image segmentation for mr brain tumor detection using machine learning: a review. *IEEE Reviews in Biomedical Engineering*, 16:70–90, 2022.

C Szegedy. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Sampada Tavse, Vijayakumar Varadarajan, Mrinal Bachute, Shilpa Gite, and Ketan Kotecha. A systematic literature review on applications of gan-synthesized images for brain mri. *Future Internet*, 14(12):351, 2022.

Zixuan Teng, Lan Li, Ziqing Xin, Dehui Xiang, Jiang Huang, Hailing Zhou, Fei Shi, Weifang Zhu, Jing Cai, Tao Peng, et al. A literature review of artificial intelligence (ai) for medical image segmentation: from ai and explainable ai to trustworthy ai. *Quantitative Imaging in Medicine and Surgery*, 14(12):9620, 2024.

Ankita Tiwari, Sampada Tavse, Mrinal Bachute, and Abhishek Bhola. A review of gan-synthesized brain mr image applications. *Revolutionizing AI with Brain-Inspired Technology: Neuromorphic Computing*, pp. 99–154, 2025.

Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data poisoning attacks against federated learning systems. In *Computer security–ESORICs 2020: 25th European symposium on research in computer security, ESORICs 2020, guildford, UK, September 14–18, 2020, proceedings, part i 25*, pp. 480–501. Springer, 2020.

Anjali Wadhwa, Anuj Bhardwaj, and Vivek Singh Verma. A review on brain tumor segmentation of mri images. *Magnetic resonance imaging*, 61:247–259, 2019.

Jindong Wang, Haoliang Li, Haohan Wang, Sinno Jialin Pan, and Xing Xie. Trustworthy machine learning: Robustness, generalization, and interpretability. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5827–5828, 2023a.

Pengyu Wang, Qiushi Yang, Zhibin He, and Yixuan Yuan. Vision transformers in multi-modal brain tumor mri segmentation: A review. *Meta-Radiology*, pp. 100004, 2023b.

Hui Wei, Hao Tang, Xuemei Jia, Zhixiang Wang, Hanxun Yu, Zhubo Li, Shin'ichi Satoh, Luc Van Gool, and Zheng Wang. Physical adversarial attack meets computer vision: A decade survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Shaoan Xie, Yanwu Xu, Mingming Gong, and Kun Zhang. Unpaired image-to-image translation with shortest path regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10177–10187, 2023.

Fahri Anıl Yerlikaya and Şerif Bahtiyar. Data poisoning attacks against machine learning algorithms. *Expert Systems with Applications*, 208:118101, 2022.

Ye Zhang, Zhixiang Wang, Zhen Zhang, Junzhuo Liu, Ying Feng, Leonard Wee, Andre Dekker, Qiaosong Chen, and Alberto Traverso. Gan-based one dimensional medical data augmentation. *Soft Computing*, 27(15):10481–10491, 2023.

Yujia Zheng, Tianhao Li, Weizhi Ma, Jiaxiang Zheng, Zhengping Li, and Lijun Wang. 5-2: Unveiling privacy challenges: Big data-driven digital twins in smart city applications. In *SID Symposium Digest of Technical Papers*, volume 55, pp. 49–52. Wiley Online Library, 2024.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A SYNTHETIC MRI RESULTS

We use a generative model (Xie et al., 2023) to synthesize fake MRI from real CT. Figure 4 presents comparative case studies of fake MRI and real MRI.

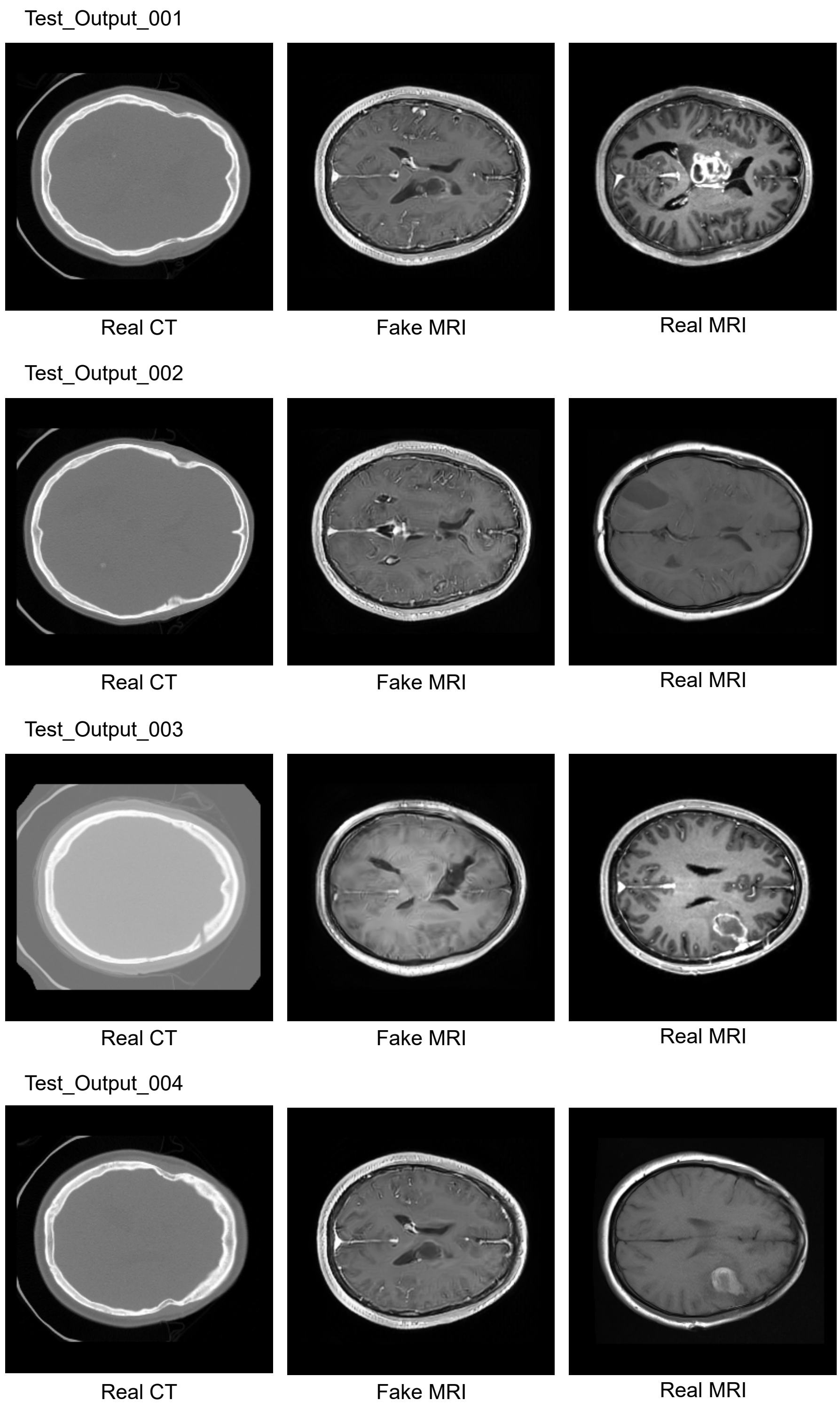

Figure 4: Case studies of synthetic MRI from real CT.