
Adversarial Robustness Analysis of Vision-Language Models in Medical Image Segmentation

Anjila Budathoki, Manish Dhakal

Department of Computer Science
Georgia State University
Atlanta, GA

Abstract

Adversarial attacks have been fairly explored for computer vision and vision-language models. However, the avenue of adversarial attack for the vision language segmentation models (VLSMs) is still under-explored, especially for medical image analysis. Thus, we have investigated the robustness of VLSMs against adversarial attacks for 2D medical images with different modalities with radiology, photography, and endoscopy. The main idea of this project was to assess the robustness of the fine-tuned VLSMs specially in the medical domain setting to address the high risk scenario. First, we have fine-tuned pre-trained VLSMs for medical image segmentation with adapters. Then, we have employed adversarial attacks—projected gradient descent (PGD) and fast gradient sign method (FGSM)—on that fine-tuned model to determine its robustness against adversaries. We have reported models’ performance decline to analyze the adversaries’ impact. The results exhibit significant drops in the DSC and IoU scores after the introduction of these adversaries. Furthermore, we also explored universal perturbation but were not able to find for the medical images.¹

1 Introduction

Artificial Intelligence (AI), especially deep neural networks, is rapidly becoming a pervasive and integral part of everyday applications, including conversational interfaces, decision support systems, and key sectors like education, healthcare, and finance [3, 4, 31]. Among these, healthcare stands out as a domain that extensively benefits from AI, spanning applications such as disease diagnosis, monitoring of various health conditions, genetic analysis, and medical image interpretation [27, 25, 2, 28, 24, 1]. In particular, medical image analysis has seen significant advancements due to deep learning, which has enabled the development of effective assistive diagnostic tools [29, 26, 12]. Given the high stakes of medical decision-making, it is essential that these models demonstrate robustness and reliability—especially since a single false negative in diagnosis can have fatal consequences [22].

Adversarial attacks apply hardly perceptible data perturbation to exploit the blind spots of the trained models, causing the models to maximize the prediction error [23]. These perturbations are not random noise, but calculated modifications to mislead the models. The attacks have been studied within the domain of vision-language models (VLMs) [6]. However, there are none to test the robustness of the vision-language segmentation models (VLSMs). VLSMs are trained to achieve the correct segmentation from images with guidance via text prompts [16, 30].

Considering the criticality of the medical image analysis and the fooling capability of adversarial attacks, there is a need to make the medical VLSMs robust against the adversarial attacks. In this research work, as a first step to fill this gap, we study the effects of adversarial images on trained

¹<https://github.com/anjilab/secure-private-ai>

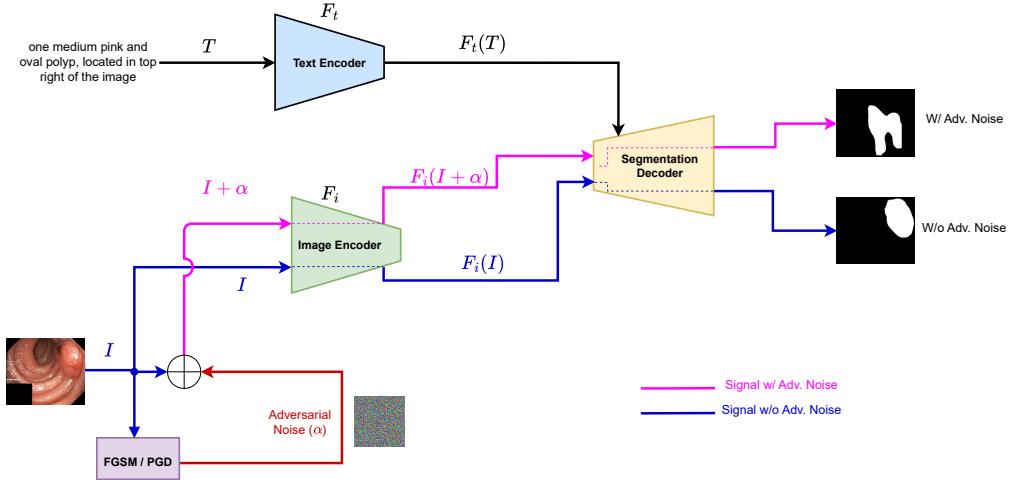


Figure 1: The overall methods of adversarial attack. F_i and F_t are fine-tuned image and text encoders, respectively. By fusing input images with the adversarial noise α generated from FGSM/PGD methods, we observe more inaccurate segmentation mask.

VLSMs. To validate this effect, we experiment with the multiple modalities of images: endoscopic, radiographic, and photographic images. The method of this paper can be broken into two major stages: (1) training a VLSM to segment target anatomical regions [5, 19] and (2) introducing adversarial attacks (PGD [18] and FGSM [8]). The comparative analysis of metrics in the presence and absence of adversarial noise exhibit the vulnerability of the VLSMs models.

1.1 Vision-Language Segmentation Models

Vision Language Segmentation Models (VLSMs) are the models trained to segment an image with guidance from the text prompts [16, 30, 14, 33]. The general approach in VLSMs training is to use two different branches of encoders to represent the text and image inputs, and the representations are passed to segmentation mask decoder. For our project, we have picked CLIPSeg [16] that has a trained decoder to segment target masks. CLIPSeg uses features from transformers-based image and text encoders of CLIP [20].

1.2 Adversarial Attacks

Adversarial attack introduce the calculated perturbation to the original image such that the model can be fooled to misguide the predictions. The impact of adversarial attacks in high-stakes domains must be addressed, as they can have serious consequences. [7] conducts an experiment across three clinical domains and were successful in both white-box and black box attacks. Previous research [17] indicates that deep neural network (DNN) models for medical images exhibit greater vulnerability compared to those for natural, non-medical images. Adversarial attacks have been extensively studied in the context of vision-language models (VLMs) [6, 21, 32]. In this study, we will further investigate the specific task of medical image segmentation.

2 Methods and Implementation

An encoder-decoder pretrained model for vision-language segmentation task is fine-tuned with a smaller training set comprising of the triplets: $D = \{(v_i, l_i, m_i)\}_{i=1}^S$. Here, S is the number of training samples, v_i , l_i , and m_i represent the image input, text prompt, and target mask of the i^{th} data point, respectively. The input images are RGB images and targets are their corresponding binary masks, i.e., $v_i \in \mathbb{R}^{H \times W \times 3}$, and $m_i \in \{0, 1\}^{H \times W}$, respectively.

The overall methodology of this research work can be divided into two major stages: fine-tuning pre-trained VLSMs for medical anatomy segmentation (section 2.1) and introduction of PGD and FGSM as adversarial attacks on those models section 2.2.

2.1 VLSM Fine-tuning

CLIPSeg [16] provides a pretrained end-to-end VLSMs. This model was trained to segment commonly seen objects—such as *dog*, *house*, *cup*, etc.—from the natural images in the real world. Without any modification to the pretrained model, it gives inferior performance when tested directly with task-specific applications like medical image segmentation. Thus, it needs to be fine-tuned. Since the encoders of CLIPSeg are large in size, fine-tuning the entire model is expensive. So, we resort to fine-tuning lightweight adapters [10] embedded within the encoders

Following the convention and training mechanism provided by VLSM-Adapter [5], we fine-tune pretrained CLIPSeg model with the addition of adapter blocks [10]. Even though [5] has provided different positioning of adapters in the image and text encoders, we use their optimal variant, *VL-Adapter*. In this variant, adapters are introduced to both image and text encoders shown in the fig. 1.

2.2 Adversarial Attacks

Our assumption for the threat model is we have access to the gradients i.e white-box attack. For both of the methods of attacks, we will perturb the input image modality as in fig. 1. Small changes in input can significantly fool state-of-the-art networks. In their work, [18] explored the impact of network capacity on adversarial attacks. Over time, models have grown to billions of parameters, and in this study, we investigate how fine-tuned, pre-trained vision-language models (VLSMs) perform against two common adversarial attack methods:

2.2.1 Fast Gradient Sign Method (FGSM)

FGSM [8] is one of the techniques for generating adversarial examples that are L_∞ bounded. To generate adversarial examples with FGSM, we compute the gradient of the loss function with respect to the input x . This gradient shows the sensitivity of the model’s loss towards changes in the input. Signs of these gradients represent the direction of perturbation that ensures maximum increase in model’s loss. ϵ scales the perturbation to control the attack. The steps can be formulated as:

$$x_* = x + \epsilon \cdot sign(\nabla_x \mathcal{L}(\theta, x, y)), \quad (1)$$

Here, x_* is the perturbed input, $\nabla_x \mathcal{L}(\theta, x, y)$ is the gradient of loss with respect to inputs, ϵ is the small constant that controls the magnitude of the perturbation and $sign(\cdot)$ is the function that gives signs of the tensor.

2.2.2 Projected Gradient Descent (PGD)

PGD [18] is an iterative extension of FGSM. In the initial iteration, the adversarial input $x_*^{(0)}$ is given as x or a slight noisy x . At each iteration, the perturbed input gets updated as:

$$x_*^{(t+1)} = clamp(x_*^{(t)} + \alpha \cdot sign(\nabla_{x_*^{(t)}} \mathcal{L}(\theta, x_*^{(t)}, y), x - \epsilon, x + \epsilon), \quad (2)$$

where α is the scaling factor of perturbation, $clamp(a, b, c)$ clamps a with the range of $[b, c]$ input within $[x - \epsilon, x + \epsilon]$, and all of the remaining symbols have similar meaning as in eq. (1).

The iteration is run for T steps and the final refined perturbed input is $x_* = x_*^{(T)}$. T is a hyperparameter, which is 40 in our experiment.

2.3 Experimentations

2.4 Implementation Setup

The training and inference of the VLSM and adversarial attack methods are executed in an NVIDIA RTX 4090. We use floating-point-16 mixed-precision training with a batch size of 32. The models are

optimized with AdamW [15] with a weight decay of $1e - 3$. The learning rate has a linear function to warmup for the first 20 epochs to reach $1e - 3$; after 20 epochs, the learning rate decays with a cosine decay function for the next 180 epochs to reach $1e - 5$. We combined dice and binary cross-entropy losses for the objective function, as shown by:

$$\mathcal{L} = \lambda_d \cdot \mathcal{L}_{Dice} + \lambda_{ce} \cdot \mathcal{L}_{CE}, \quad (3)$$

where λ_d and λ_{ce} are hyperparameters; we chose their values for our experiments as $\lambda_d = 1.5$ and $\lambda_{ce} = 1$.

The bottleneck layer’s dimension of the adapter is 64. Adapters are added to both image and text encoder branches of the model. The image has been resized to 352×352 for batch processing, and the context size for text input is 77.

During the noise injection, different scales $\epsilon \in \{0.01, 0.03, 0.1, 0.5\}$ are used to determine the amount of perturbation to be introduced in the input images.

2.5 Datasets

Poudel et al. [19] published a variety of language prompts grounded to the target object for medical image segmentation. We have selectively sampled a few datasets that represent a wider range of modalities within radiology and non-radiology medical images. We have worked with Kvasir-SEG [11] for endoscopic images, ISIC-16 [9] for photographic images, and CAMUS [13] for radiographic images.

2.6 Evaluation metrics

We have used two evaluation metrics popular in medical image segmentation, dice score (DSC) and intersection-over-union (IoU) as:

$$DSC = \frac{2 * (y_{pred} \cap y_{true})}{y_{pred} + y_{true}}, \quad (4)$$

$$IoU = \frac{y_{pred} \cap y_{true}}{y_{pred} \cup y_{true}}, \quad (5)$$

where y_{pred} and y_{true} are predicted and targeted binary masks. For a successful adversarial attack, we compare the metrics before and after the attack.

3 Results

Table 1 compares the performance of two different attack methods: FGSM and PGD in terms of two evaluation metrics, dice score and intersection-over-union, measured under varying perturbation levels for the given three datasets. We have chosen four perturbation levels for the study. The table shows that the attack was successful, as evidenced by the decreasing scores with increasing perturbation levels across all the datasets.

Figure 2 illustrates the images generated using FGSM for the Kvasir-SEG dataset. As the perturbation size increases, the added perturbations become more noticeable, with higher visibility at $\epsilon = 0.5$ compared to $\epsilon = 0.01$. For additional generated adversarial images for the other two datasets, please refer to A.1

4 Discussion

Radiological (Grayscale) images are more vulnerable. Among the evaluated datasets, CAMUS experienced a more pronounced drop in DSC and IoU compared to the others under the same hyperparameter settings. A possible explanation for this discrepancy lies in the nature of the input images. While other datasets use three-channel RGB images, CAMUS consists of single-channel images representing pixel brightness.

Dataset	Adversarial Attack	ϵ	DSC% \uparrow	IoU % \uparrow
	Original (W/o attack)	-	88.83	82.72
Kvasir-SEG [11]	FGSM	0.01	75.08	64.31
		0.03	67.57	55.91
		0.1	58.36	45.97
		0.5	51.81	39.49
	PGD	0.01	71.30	61.84
		0.03	79.17	71.25
		0.1	47.78	37.43
		0.5	37.69	27.63
	Original (W/o attack)	-	92.27	86.29
ISIC-16 [9]	FGSM	0.01	84.38	74.79
		0.03	80.23	69.02
		0.1	75.84	63.14
		0.5	74.05	61.51
	PGD	0.01	89.34	82.15
		0.03	90.36	83.71
		0.1	83.24	74.65
		0.5	82.65	73.22
	Original (W/o attack)	-	89.87	82.13
CAMUS [13]	FGSM	0.01	75.99	63.29
		0.03	73.25	60.05
		0.1	71.05	57.39
		0.5	48.17	35.12
	PGD	0.01	46.56	36.10
		0.03	34.21	24.85
		0.1	15.16	9.22
		0.5	14.47	8.6

Table 1: Comparison of dice score and intersection of union scores across two adversarial attacks: FGSM and PGD for four perturbation values in three datasets: Kvasir-SEG [11], ISIC-16 [9], CAMUS [13].

We hypothesize that in RGB images, if an adversarial attack affects one channel of a pixel, the remaining two channels can still retain information, partially compensating for the loss — assuming those channels remain unaffected. In contrast, for an attack to fully disrupt an RGB pixel, all three channels would need to be targeted simultaneously, which is less likely.

Computation vs attack success vs imperceptibility. When comparing FGSM and PGD in terms of computation, attack success, and imperceptibility, some clear trade-offs become evident. FGSM is a single-step attack that’s computationally lightweight and easy to implement. However, its success rate tends to be lower (refer to table 1) than iterative methods, and it offers limited control over how noticeable the perturbations are. At higher attack strengths, it often results in visible artifacts. On the other hand, PGD builds on FGSM by applying small, incremental perturbations over multiple steps, projecting the adversarial example back within an ϵ -ball, $[x - \epsilon, x + \epsilon]$, around the original input after each step. While this increases computational cost, it leads to much higher attack success rates and better imperceptibility (refer to fig. 2 and appendix A.1). Thanks to its gradual and controlled updates, PGD is widely regarded as one of the strongest and more imperceptible first-order attacks, often producing adversarial examples that are difficult to distinguish from the original images.

5 Conclusion and Future Work

We implemented adversarial attacks with FGSM and PGD in vision language segmentation models for medical images data. The findings suggest the attack on these models was successful with marginal decrease in dice score and intersection-over-union. This is a small stepping stone towards making VLSMs robust to such data poisoning methods.

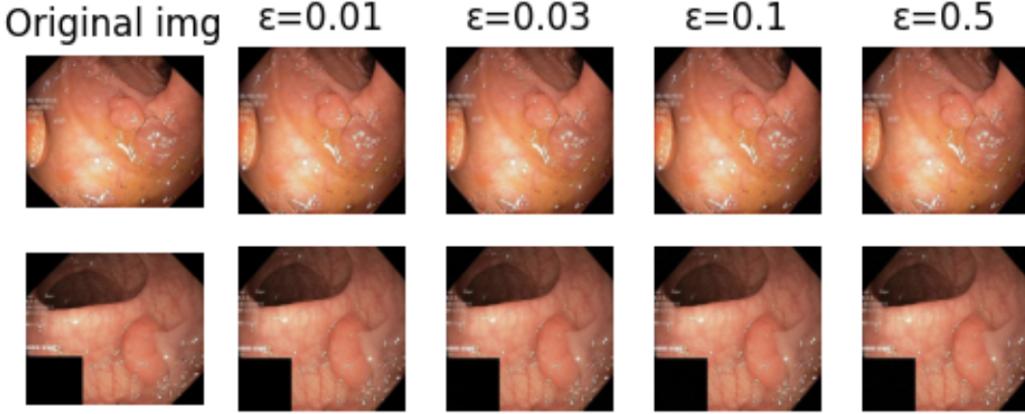


Figure 2: Comparison of original images and adversarial images generated at different perturbation levels. As the perturbation increases, the adversarial modifications become increasingly perceptible to the human eye.

In this study, we have identified that the issues of adversarial attack also persist in the VLSM domains. For future studies, we can research enabling the defense mechanisms against the attack. Also, we only have studied the white-box attacks of the models; however, we need to explore black-box attacks in which we have limited access to execution, gradients, and parameters. We have tested it only one VLSM (i.e. CLIPSeg), but the effects of adversarial attacks must be studied across other pre-existing segmentation models. In the future, we aim to implement universal adversarial perturbation method (i.e., one adversary to attack all images in the dataset).

References

- [1] Rabin Adhikari, Manish Dhakal, Safal Thapaliya, Kanchan Poudel, Prasiddha Bhandari, and Bishesh Khanal. Synthetic boost: leveraging synthetic data for enhanced vision-language segmentation in echocardiography. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 89–99. Springer, 2023. 1
- [2] Gopi Battineni, Getu Gamo Sagaro, Nalini Chinatalapudi, and Francesco Amenta. Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of personalized medicine*, 10(2):21, 2020. 1
- [3] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. Enhancing ai-assisted group decision making through llm-powered devil’s advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 103–119, 2024. 1
- [4] I de Zarzà, J de Curtò, Gemma Roig, and Carlos T Calafate. Optimized financial planning: integrating individual and cooperative budgeting models with llm recommendations. *AI*, 5(1):91–114, 2023. 1
- [5] Manish Dhakal, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Vlsm-adapter: Fine-tuning vision-language segmentation efficiently with lightweight blocks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 712–722. Springer, 2024. 1, 2.1
- [6] Junhao Dong, Junxi Chen, Xiaohua Xie, Jianhuang Lai, and Hao Chen. Adversarial attack and defense for medical image analysis: Methods and applications. *arXiv preprint arXiv:2303.14133*, 2023. 1, 1.2
- [7] Samuel G Finlayson, Hyung Won Chung, Isaac S Kohane, and Andrew L Beam. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*, 2018. 1.2
- [8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2.2.1

- [9] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 2.5, 3, 1
- [10] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 2.1
- [11] Debesh Jha, Pia H Smedsrød, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, pages 451–462. Springer, 2020. 2.5, 3, 1
- [12] Naresh Kumar Kar, S Jana, Abdur Rahman, Patil Rahul Ashokrao, R Alarmelu Mangai, et al. Automated intracranial hemorrhage detection using deep learning in medical image analysis. In *2024 International Conference on Data Science and Network Security (ICDSNS)*, pages 1–6. IEEE, 2024. 1
- [13] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019. 2.5, 3, 1
- [14] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023. 1.1
- [15] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2.4
- [16] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022. 1, 1.1, 2.1
- [17] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021. 1.2
- [18] Aleksander Madry. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2.2, 2.2.2
- [19] Kanchan Poudel, Manish Dhakal, Prasiddha Bhandari, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Exploring transfer learning in medical image segmentation using vision-language models. *arXiv preprint arXiv:2308.07706*, 2023. 1, 2.5
- [20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1.1
- [21] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations*, 2023. 1.2
- [22] Vera Sorin, Shelly Soffer, Benjamin S Glicksberg, Yiftach Barash, Eli Konen, and Eyal Klang. Adversarial attacks in radiology—a systematic review. *European journal of radiology*, 167:111085, 2023. 1
- [23] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014. 1

- [24] Cynthia Y Tang, Cheng Gao, Kritika Prasai, Tao Li, Shreya Dash, Jane A McElroy, Jun Hang, and Xiu-Feng Wan. Prediction models for covid-19 disease outcomes. *Emerging Microbes & Infections*, 13(1):2361791, 2024. 1
- [25] Muhammad Ali Javed Tengnah, Raginee Sooklall, and Soulakshmee Devi Nagowah. A predictive model for hypertension diagnosis using machine learning techniques. In *Telemedicine technologies*, pages 139–152. Elsevier, 2019. 1
- [26] Bishal Thapaliya, Vince D Calhoun, and Jingyu Liu. Environmental and genome-wide association study on children anxiety and depression. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2330–2337. IEEE, 2021. 1
- [27] Bishal Thapaliya, Bhaskar Ray, Britny Farahdel, Pranav Suresh, Ram Sapkota, Bharath Holla, Jayant Mahadevan, Jiayu Chen, Nilakshi Vaidya, Nora Irma Perrone-Bizzozero, et al. Cross-continental environmental and genome-wide association study on children and adolescent anxiety and depression. *Frontiers in Psychiatry*, 15:1384298, 2024. 1
- [28] Bishal Thapaliya, Zundong Wu, Ram Sapkota, Bhaskar Ray, Pranav Suresh, Santosh Ghimire, Vince Calhoun, and Jingyu Liu. Graph-based deep learning models in the prediction of early-stage alzheimers. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE, 2024. 1
- [29] Jiaji Wang, Shuihua Wang, and Yudong Zhang. Deep learning on medical image analysis. *CAAI Transactions on Intelligence Technology*, 10(1):1–35, 2025. 1
- [30] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 1, 1.1
- [31] Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. Evaluating reading comprehension exercises generated by llms: A showcase of chatgpt in education applications. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 610–625, 2023. 1
- [32] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024. 1.2
- [33] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11175–11185, June 2023. 1.1

A Appendix

A.1 Adversarial images generated

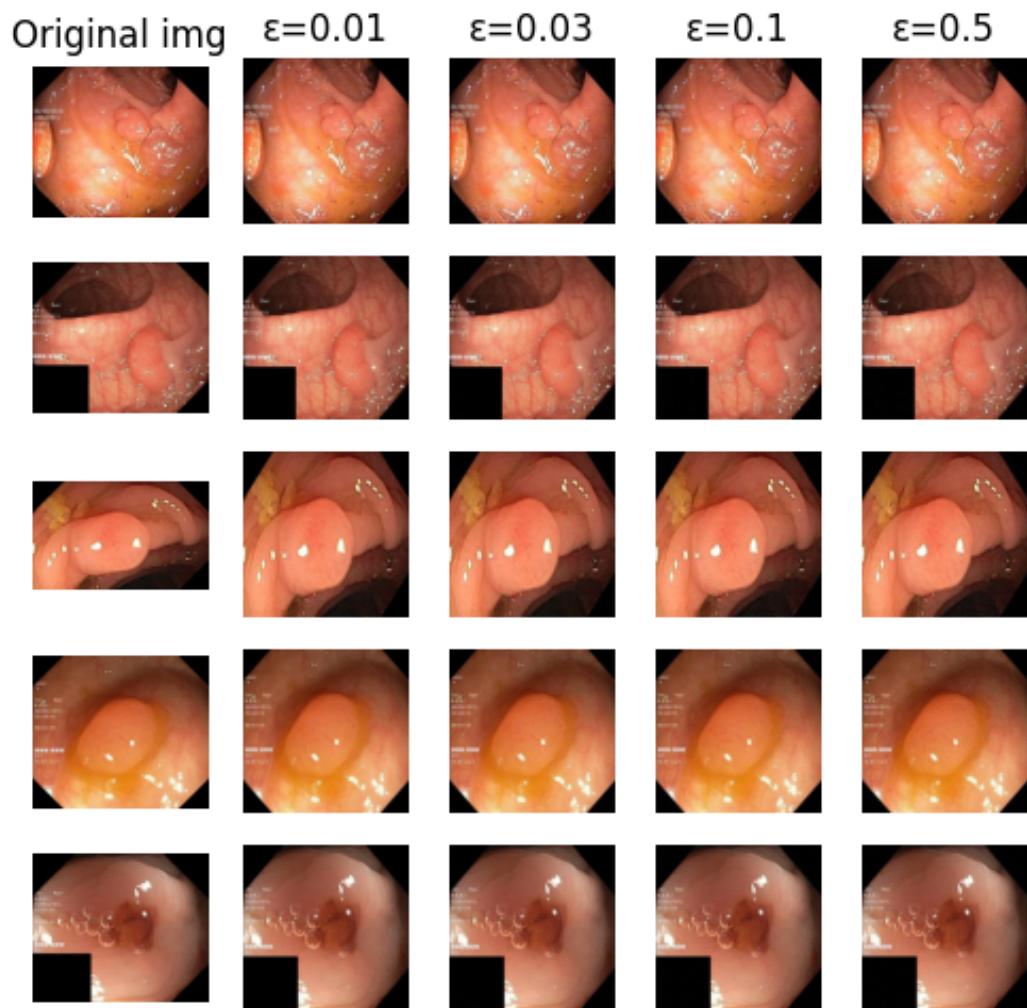


Figure 3: Comparison of original images and adversarial images generated at different perturbation levels on Kvasir dataset using PGD. As the perturbation increases, the adversarial modifications become increasingly perceptible to the human eye.

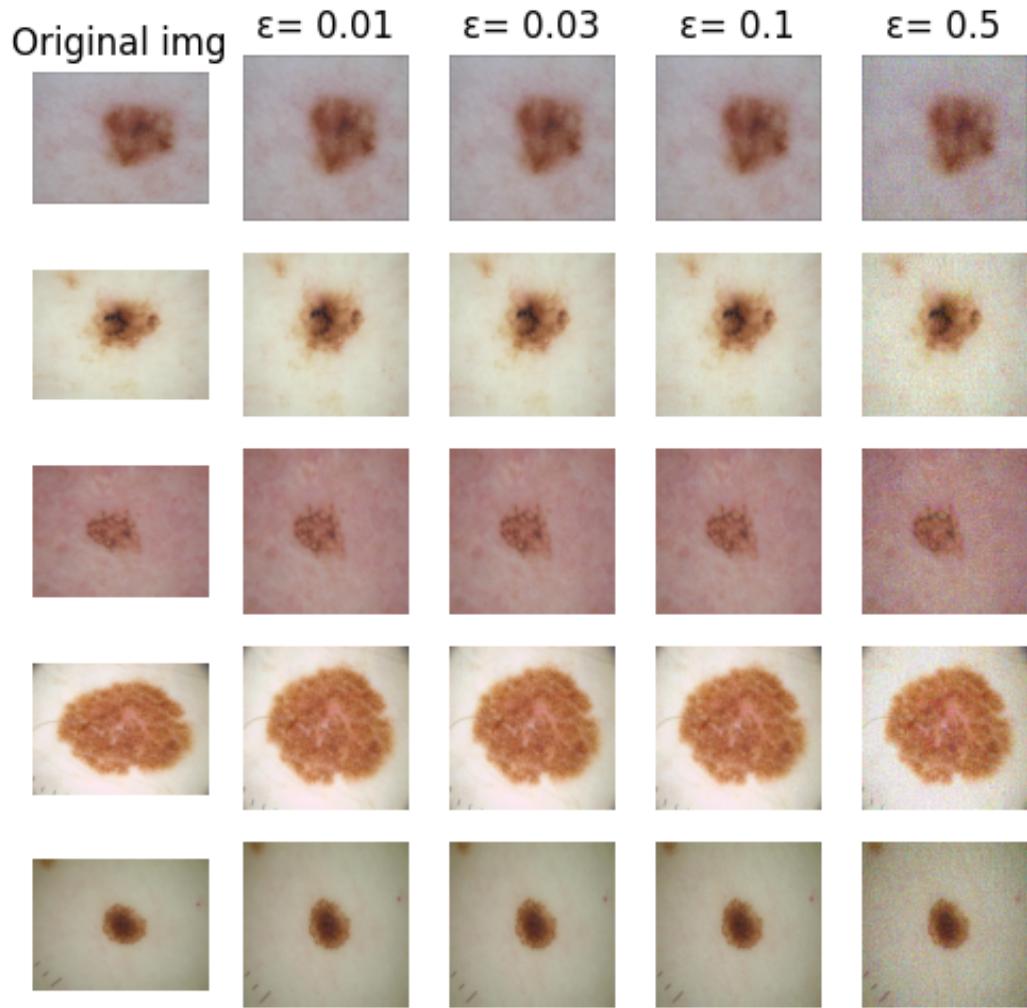


Figure 4: Comparison of original images and adversarial images generated at different perturbation levels on ISIC-16 dataset using FGSM. As the perturbation increases, the adversarial modifications become increasingly perceptible to the human eye.

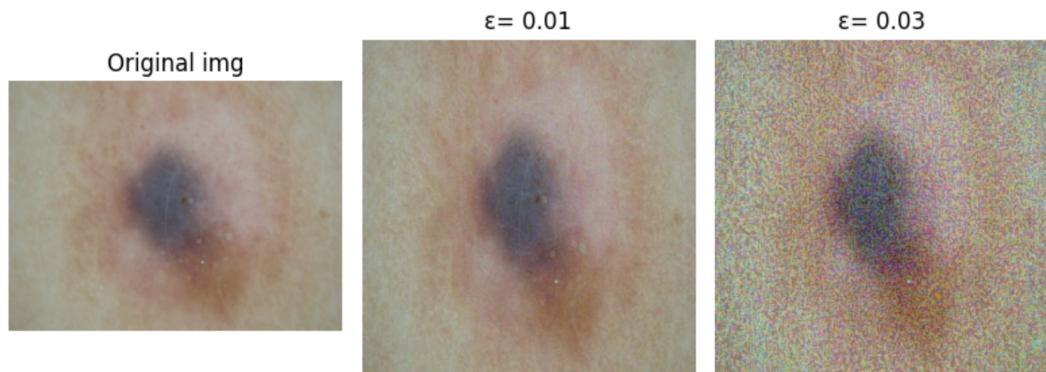


Figure 5: The provided image is a zoomed-in version highlighting the differences between the adversarial and original images. It offers a clearer comparison between the original and perturbed images generated via FGSM. This pattern is consistent across all other samples.

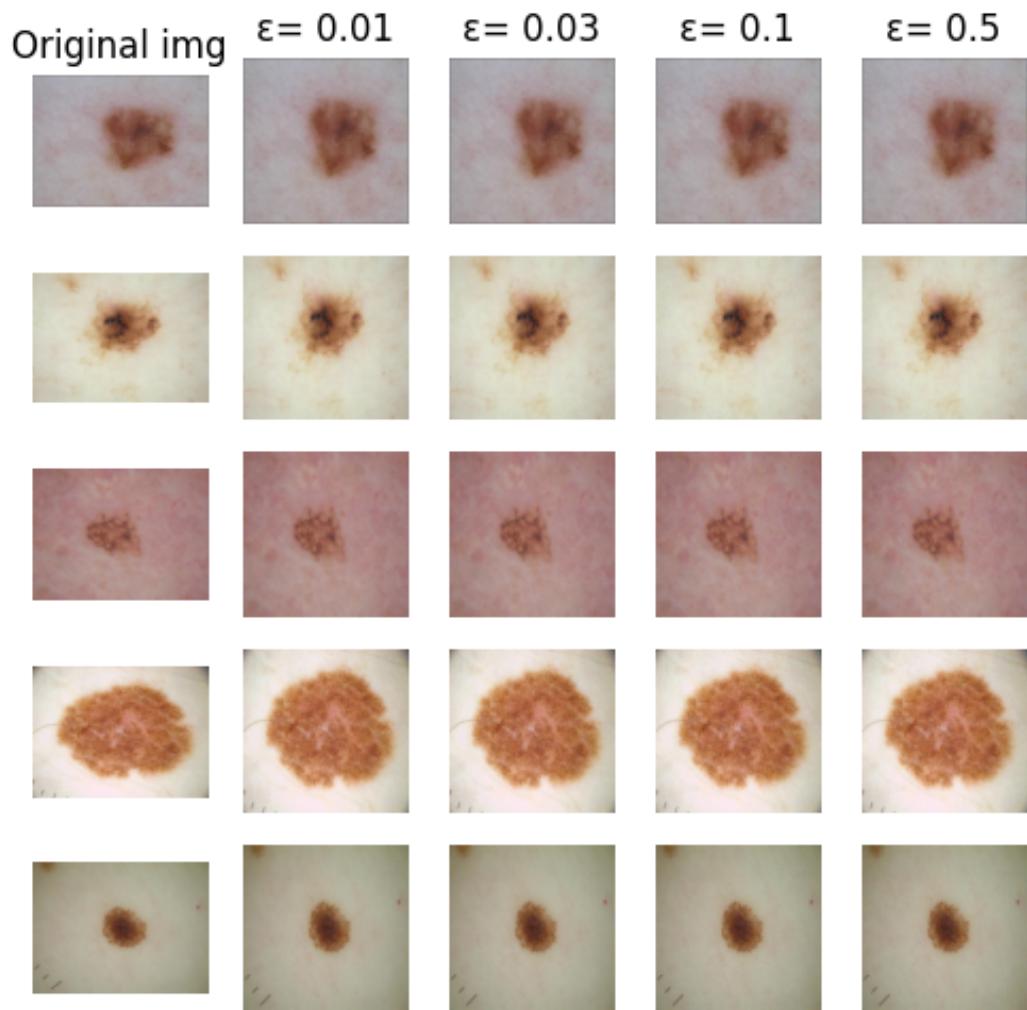


Figure 6: Comparison of original images and adversarial images generated at different perturbation levels on ISIC-16 dataset using PGD. As the perturbation increases, the adversarial modifications become increasingly perceptible to the human eye.

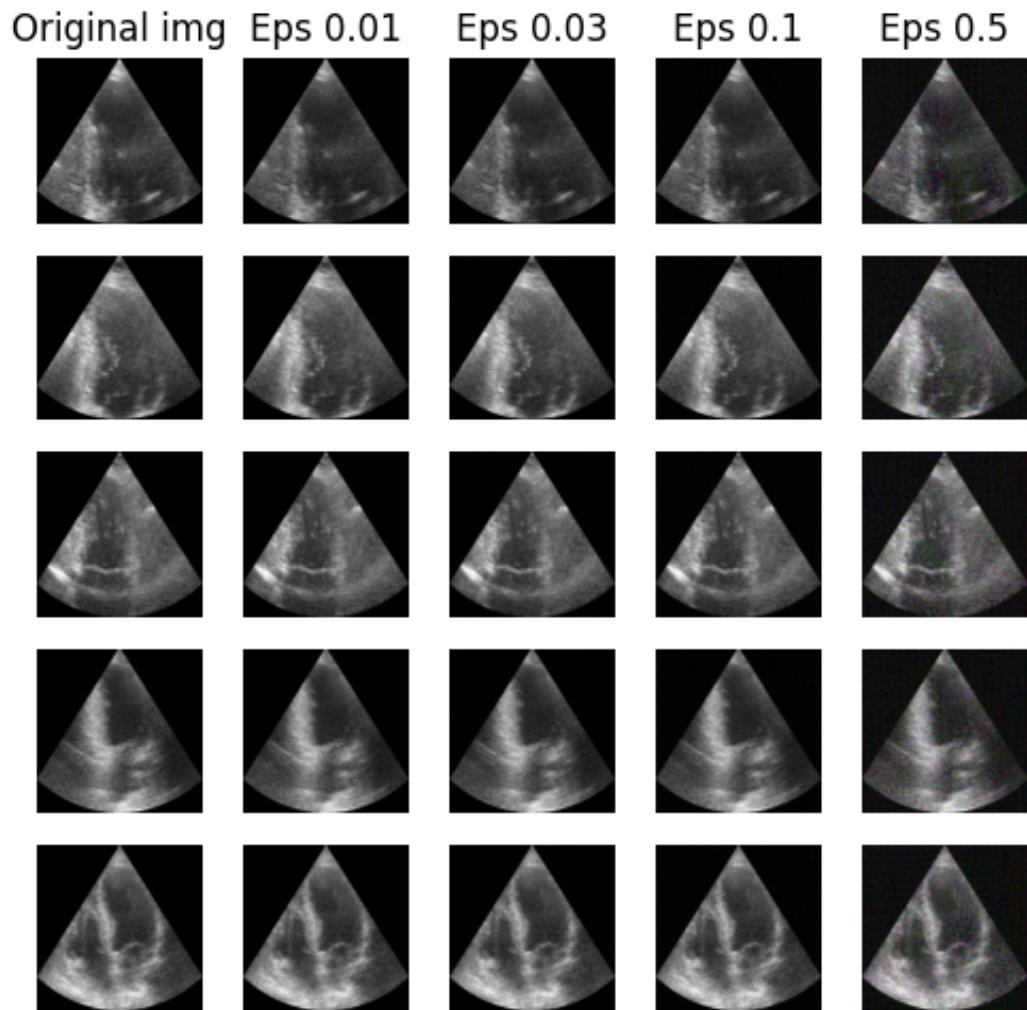


Figure 7: Comparison of original images and adversarial images generated at different perturbation levels on CAMUS dataset using FGSM. As the perturbation increases, the adversarial modifications become increasingly perceptible to the human eye.

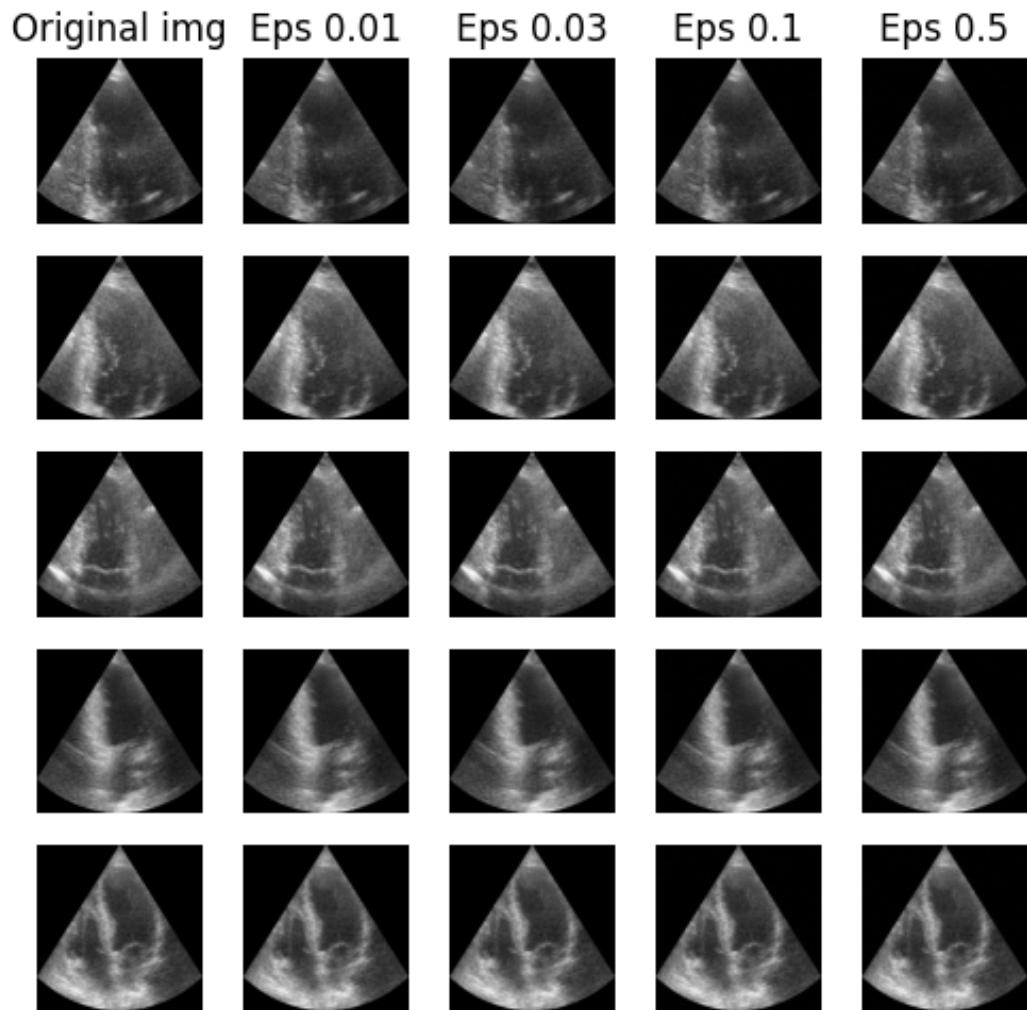


Figure 8: Comparison of original images and adversarial images generated at different perturbation levels on CAMUS dataset using PGD. As the perturbation increases, the adversarial modifications become increasingly perceptible to the human eye.