DEMONSTRATION OF AN ADVERSARIAL ATTACK AGAINST A MULTIMODAL VISION LANGUAGE MODEL FOR PATHOLOGY IMAGING

Poojitha Thota[⋆,1] *Jai Prakash Veerla*[⋆,1,3] *Partha Sai Guttikonda*[1,3]
Mohammad S. Nasr[1,3] *Shirin Nilizadeh*[†,1,3] *Jacob M. Luber*[†,1,2,3]

[1] Department of Computer Science and Engineering, University of Texas at Arlington
[2] Department of Bioengineering, University of Texas at Arlington
[3] Multi-Interprofessional Center for Health Informatics, University of Texas at Arlington

ABSTRACT

In the context of medical artificial intelligence, this study explores the vulnerabilities of the Pathology Language-Image Pretraining (PLIP) model, a Vision Language Foundation model, under targeted attacks. Leveraging the Kather Colon dataset with 7,180 H&E images across nine tissue types, our investigation employs Projected Gradient Descent (PGD) adversarial perturbation attacks to induce misclassifications intentionally. The outcomes reveal a 100% success rate in manipulating PLIP's predictions, underscoring its susceptibility to adversarial perturbations. The qualitative analysis of adversarial examples delves into the interpretability challenges, shedding light on nuanced changes in predictions induced by adversarial manipulations. These findings contribute crucial insights into the interpretability, domain adaptation, and trustworthiness of Vision Language Models in medical imaging. The study emphasizes the pressing need for robust defenses to ensure the reliability of AI models. The source codes for this experiment can be found at https://github.com/jaiprakash1824/VLM_Adv_Attack.

Index Terms— Adversarial Attacks, Histopathology Data, Vision Language Foundation Models, Pathology, AI, Robustness, Trustworthiness, Medical Image Analysis

1. INTRODUCTION

The incorporation of artificial intelligence (AI) into the field of medical imaging and pathology has experienced notable advancements, leading to substantial progress in the areas of diagnoses and analysis [1, 2]. Furthermore, the utilization of AI in the field of pathology goes beyond traditional imaging methods and encompasses the intricacies associated with Hematoxylin and Eosin (H&E) staining [3]. This integration allows for a thorough comprehension of tissue structures and the morphology of cells.

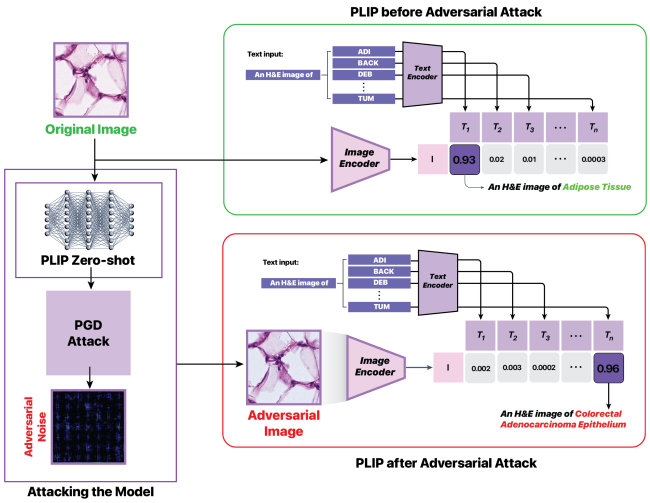

Fig. 1. Attack Overview

PLIP, a vision language model for pathology, is an AI framework that operates in several dimensions, effectively managing the complex interplay between visual and textual data [4]. It leverages the collective knowledge found on platforms such as medical Twitter to overcome the limitations posed by a scarcity of annotated medical images. This significant advancement not only tackles the issue of limited data availability but also enables pioneering research in the field of pathology analysis. These models have exhibited exceptional performance in zero-shot classification, representing cutting-edge capabilities in this domain.

With the increasing integration of AI models like PLIP into pathology practice, the potential threat of adversarial attacks becomes a significant concern. The reliability of AI systems in medical imaging can be compromised by many techniques, such as Fast Gradient Sign Method (FGSM) at-

⋆Co-Authors - Equal contribution; Order of authors is alphabetical and both first authors have the right to list their name first on their respective CVs.
†Responsible authors. Email: jacob.luber@uta.edu, shirin.nilizadeh@uta.edu

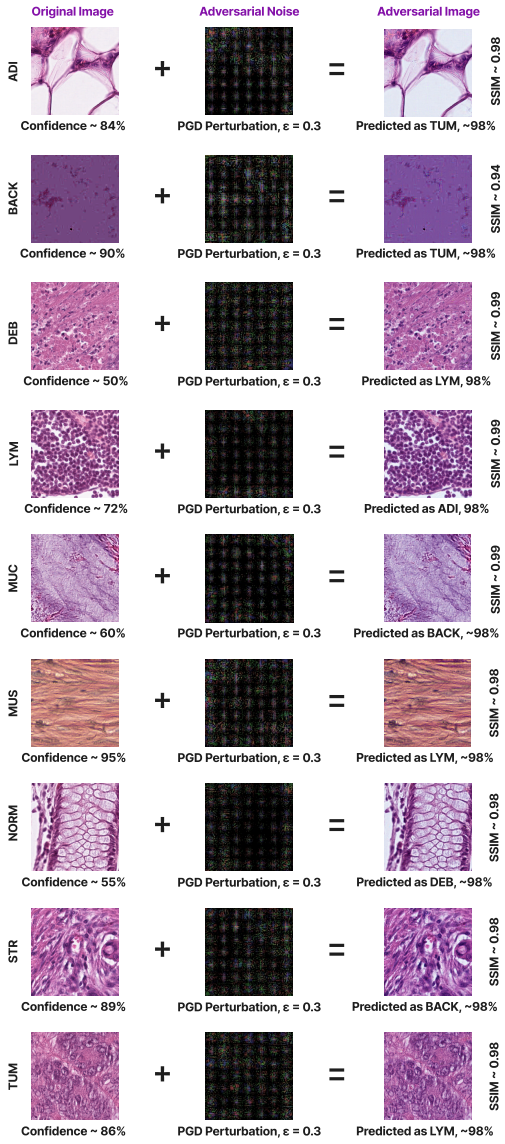

Fig. 2. Original H&E + Perturbation = Adversarial H&E

tack [5], Carlini & Wagner [6], and projected gradient descent (PGD) attack [7]. These attacks have the potential to pose significant challenges to the integrity and accuracy of AI systems in this domain.

Interpretability and trustworthiness are very important to the field of pathology. The lack of transparency and comprehensibility in the decision-making process of numerous machine learning models gives rise to issues regarding their black-box character and the capacity to understand the underlying reasoning behind their forecasts.

In this domain, the ability to make informed decisions relies heavily on the comprehensibility of the outputs generated by AI systems. By conducting a thorough investigation of attacks and emphasizing the need for interpretability and trustworthiness, our objective is to significantly contribute to the academic discussion on AI's impact on pathology analysis. We demonstrate this through the first successful adversarial attack against a pathology vision language model.

2. METHODS

2.1. Threat Model

Our examination of the threat model centers on the reliability of Vision Language Models (VLMs), specifically PLIP. A hypothetical situation where such an adversarial attack might manifest would be a payer altering diagnostic results with the intention of obtaining financial benefits, such as denying an insurance claim.

2.2. Vision Language Models - PLIP

Our investigation focuses on the methodological aspects of the PLIP model, including examining its architectural intricacies. PLIP has been chosen based on its exceptional ability to include both visual and textual data. The multimodality inherent to vision language models is achieved by leveraging medical Twitter to carefully select and organize the OpenPath dataset. This dataset is a substantial resource consisting of 208,414 pathology images accompanied by descriptions in natural language. This model has been rigorously evaluated on popular datasets, such as KatherColon [8], PanNuke [9], DigestPath [10] datasets for zero-shot image classification.

2.3. Dataset

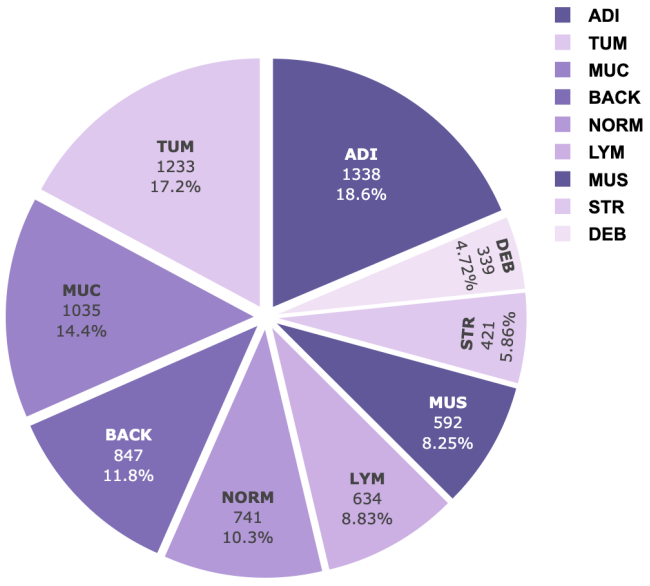

Fig. 3. PieChart of Kather Colon Dataset

Our study relies on the Kather Colon dataset, a carefully selected 7,180 image patches from 50 colorectal adenocar-

cinoma patients. The dataset used in this study consists of photographs with dimensions of 224x224 pixels and a resolution of 0.5 megapixels per pixel (MPP). It serves not only as a testing platform but also as a validation set with clinical relevance, hence assuring the practicality and application of our research findings in real-world scenarios. Fig.3, is a piechart representing 9 tissue types with the number of images in each tissue type.

2.4. Adversarial Attack - Projected Gradient Descent

We chose to perform an adversarial attack on PLIP by performing the PGD method. The choice of PGD is supported by its efficacy as an iterative optimization method, enabling the creation of perturbations that are both effective and subtle. This particular attribute contributes to our objective of methodically examining and comprehending the vulnerabilities of PLIP under controlled hostile circumstances. This enables a thorough evaluation of the resilience of PLIP against sophisticated attacks.

As shown in Fig. 1, we take the original H&E image and perform the PGD attack by iterating through several steps to achieve optimal perturbation. We perform this attack based on two main objectives, which are targeted misclassification and a high Structural Similarity Index Measure (SSIM) score.

3. EXPERIMENTS AND RESULTS

3.1. Evaluation

In order to provide a clearer understanding of the influence of adversarial attacks on the predictions made by PLIP, we utilized heat maps as a visual tool to depict the model's label predictions both prior to and subsequent to the attacks, see Fig. 5. A continuous pattern of accurate label predictions is observed in the top heat map obtained from the original photos. However, the observed pattern is significantly disturbed in the associated adversarial heat maps at the bottom of Fig. 5. This investigation highlights the difficulty presented by adversarial manipulations, placing emphasis on the significance of interpretability in medical artificial intelligence (AI) models.

To strengthen our assessment, we calculated the SSIM ratings for both the unaltered and manipulated images. Notably, it was noted that the majority of SSIM values are above 90%, suggesting a significant level of resemblance between the unaltered and modified photos. The targeted PGD attacks aimed at inducing misclassifications within PLIP's predictions for specific tissue types were remarkably successful, yielding a 100% attack success rate. From Fig. 6, we observe that we have reached a 100% attack success rate for all the tissue types after 10 steps to achieve optimal perturbation, demonstrating the vulnerability of the vision language model.

Moreover, we visualized the attention patterns of the PLIP model before and after subjecting it to a PGD attack, i.e. both

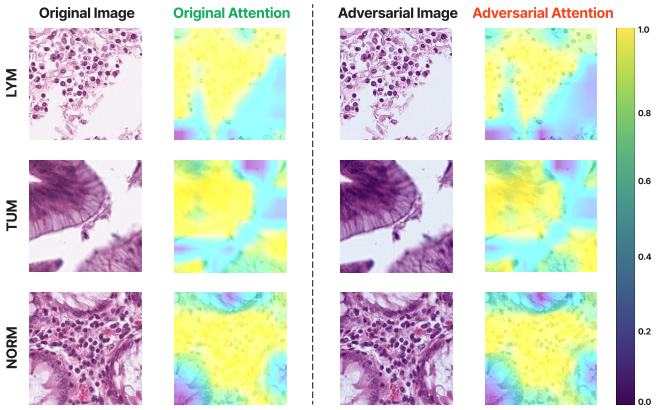

Fig. 4. Visualization of Attention before and after PGD attack on PLIP

on the original and perturbed images. This examination provided valuable insights into the specific areas where the model focuses its attention and identifies regions crucial for classification [11].

Upon inspecting the tissue images labeled LYM and NORM in Fig. 4, we observed distinct attention patterns. In the original attention distribution (left side of Fig. 4), the model exhibited heightened attention (indicated by the predominance of yellow) on various cell structures. However, following the PGD attack, represented by the adversarial attention distribution (right side of Fig. 4), we noticed a noticeable shift or reduction in attention towards these cell structures. In particular, the attention allocated to these structures decreased (illustrated by the prevalence of green) compared to the original attention distribution. Interestingly, the attention appeared to redistribute towards the surrounding areas of the cell structures.

This analysis underscores the dynamic nature of the model's attention mechanism and highlights its susceptibility to adversarial perturbations. By discerning these changes in attention allocation, we gain valuable insights into the model's decision-making process and its sensitivity to external manipulations.

3.2. Targeted Misclassification

In the context of our targeted PGD adversarial attacks on the PLIP model using the Kather Colon dataset, our objective was to intentionally induce misclassifications for specific tissue types.

These targeted misclassifications were chosen to assess the model's susceptibility to adversarial manipulations across a spectrum of tissue types, mirroring potential real-world scenarios, where misdiagnoses could have critical consequences. The successful implementation of these targeted misclassifications, which can be observed from Fig. 2, further underscores the nuanced vulnerabilities of PLIP to adversarial at-

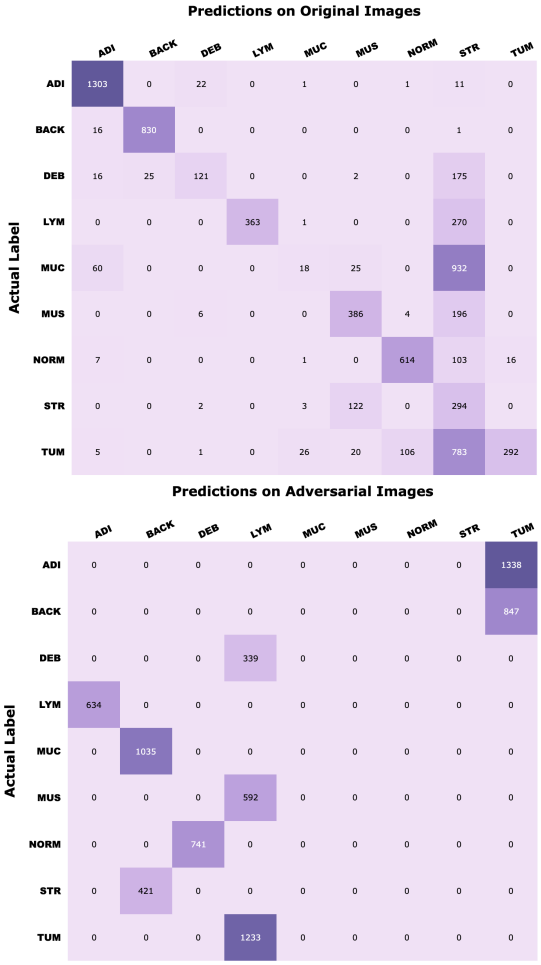

Fig. 5. Heatmaps showing the distribution of Adversarial Attacks

tacks and raises questions about the model's trustworthiness in medical imaging applications. Along with PLIP, we have also performed this targeted PGD adversarial attack on an additional VLM model, BiomedClip, in which we have observed similar vulnerabilities of misclassifications (data not shown). [12]

4. DISCUSSIONS

4.1. Possible Defenses

In light of vulnerabilities exhibited by PLIP model under targeted PGD adversarial attack, incorporating robust defense strategies becomes necessary to safeguard the model against such threats. To improve robustness, Adversarial training has been proven to be one of the most effective approaches in the image domain [13]. However, understanding that the approach of adversarial training is computationally intensive, especially in the case of VLMs, recent works have explored input pre-processing techniques such as diffusion-based purification, to improve the robustness of VLMs [14] [15]. These methods rely on generative models to purify adversarial perturbations before classification, without any need for re-training. Given the critical nature of pathology image analysis and the potential implications of adversarial attacks in medical diagnostics, VLMs can be made more resilient to adversarial attacks by exploring a combination of these advanced strategies, thereby ensuring their security and reliability in real-world applications.

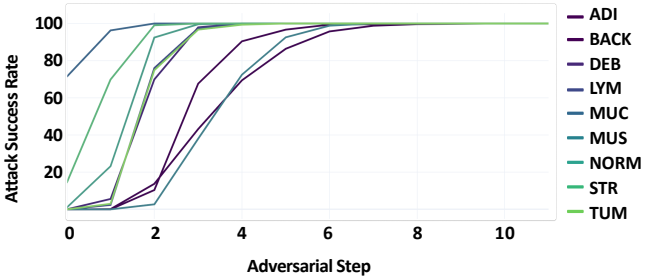

Fig. 6. Attack Success Rate per step

5. CONCLUSION AND FUTURE DIRECTIONS

In the pursuit of advancing the interpretability, domain adaptation, and trustworthiness of medical artificial intelligence, our investigation into the vulnerabilities of the PLIP model through targeted PGD adversarial attacks reveals critical insights. The 100% success rate in inducing intentional misclassifications underscores PLIP's susceptibility to adversarial manipulations, prompting a fundamental reassessment of its trustworthiness in medical imaging applications. The qualitative analysis of adversarial examples illuminates the changes in PLIP's predictions, emphasizing the need for interpretability-aware defenses to fortify models. These findings contribute to the broader discourse on the robustness of Vision Language Models in pathology analysis, guiding the development of AI models that not only exhibit high performance but also maintain reliability in the face of adversarial attacks.

6. ACKNOWLEDGMENTS

This work was supported by the University of Texas System Rising STARs Award (J.M.L) and the CPRIT First Time Faculty Award (J.M.L). Additionally, the research presented in this paper was supported by the National Science Foundation under Grant No. CNS: 2239646.

7. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access. Ethical ap-

proval was **not** required as confirmed by the license attached with the open access data.

8. REFERENCES

[1] Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood, "Ai-based pathology predicts origins for cancers of unknown primary," *Nature*, vol. 594, no. 7861, pp. 106–110, 2021.

[2] Annie Y Ng, Cary JG Oberije, Éva Ambrózay, Endre Szabó, Orsolya Serfőző, Edit Karpati, Georgia Fox, Ben Glocker, Elizabeth A Morris, Gábor Forrai, et al., "Prospective implementation of ai-assisted screen reading to improve early detection of breast cancer," *Nature Medicine*, pp. 1–6, 2023.

[3] Kevin de Haan, Yijie Zhang, Jonathan E Zuckerman, Tairan Liu, Anthony E Sisk, Miguel FP Diaz, Kuang-Yu Jen, Alexander Nobori, Sofia Liou, Sarah Zhang, et al., "Deep learning-based transformation of h&e stained tissues into special stains," *Nature communications*, vol. 12, no. 1, pp. 4884, 2021.

[4] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical twitter," *Nat Med*, vol. 29, no. 9, pp. Art. no. 9, Sep 2023.

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[6] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017, pp. 39–57.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," 2019.

[8] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al., "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS medicine*, vol. 16, no. 1, pp. e1002730, 2019.

[9] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot, "Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification," in *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*. Springer, 2019, pp. 11–19.

[10] Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, et al., "Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system," *Medical Image Analysis*, vol. 80, pp. 102485, 2022.

[11] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal, "Vl-interpret: An interactive visualization tool for interpreting vision-language transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21406–21415.

[12] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon, "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," 2024.

[13] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *International conference on machine learning*. PMLR, 2021, pp. 11492–11501.

[14] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar, "Diffusion models for adversarial purification," *arXiv preprint arXiv:2205.07460*, 2022.

[15] Changhao Shi, Chester Holtz, and Gal Mishne, "Online adversarial purification based on self-supervised learning," in *International Conference on Learning Representations*, 2021.