Adversarial Robustness in Medical Image Classification: Defending Against Attacks for Safer Healthcare AI

Bhuvan Karthik Channagiri
Northeastern University
channagiri.b@northeastern.edu

Nikita Vinod Mandal
Northeastern University
mandal.n@northeastern.edu

***Abstract*—This project investigates the robustness of medical image classifiers for lung and colon histopathological data against adversarial attacks. Using convolutional neural networks (CNNs) as a baseline, we implement adversarial attacks, including Projected Gradient Descent (PGD), Universal Adversarial Perturbations (UAP), and Patch Attacks, to evaluate vulnerabilities. To counteract these threats, we explore defense mechanisms such as adversarial pretraining and Vision Transformers (ViTs), focusing on their capacity to model global dependencies and resist perturbations.**

The study highlights the limitations of traditional CNNs in handling adversarial scenarios and demonstrates the efficacy of ViTs and adversarially robust pretraining in enhancing model safety and reliability. By combining advanced architectures and targeted defenses, the project provides valuable insights into building robust AI systems for critical applications in healthcare, emphasizing the need for secure and trustworthy medical diagnostics powered by machine learning.

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The advent of machine learning has brought transformative advancements to the healthcare industry, enabling precise, efficient, and automated diagnostic tools. Among these, medical image classification has emerged as a critical task, assisting clinicians in identifying complex patterns within histopathological data. For lung and colon cancer diagnosis, histopathology images play a pivotal role in determining cancer subtypes, aiding in timely and accurate interventions.

However, machine learning models, particularly neural networks, face a significant challenge: their vulnerability to adversarial attacks. These attacks exploit the model's susceptibility by introducing small, often imperceptible perturbations to input data, resulting in incorrect predictions. In a healthcare setting, such vulnerabilities could lead to misdiagnoses, jeopardizing patient safety and undermining trust in AI-driven tools.

To address these challenges, this project investigates the robustness of convolutional neural networks (CNNs) trained on a curated dataset of lung and colon cancer histopathology images. The dataset encompasses five distinct classes, providing a comprehensive foundation for evaluating model performance. Adversarial attack methods, including Projected Gradient Descent (PGD), Patch Attack, and Universal Adversarial Perturbations (UAP), are employed to test the model's susceptibility to pixel-wise, localized, and dataset-wide perturbations, respectively.

To counter these vulnerabilities, we introduce advanced defensive mechanisms aimed at fortifying the model's robustness:

Vision Transformers (ViTs): Leveraging their global self-attention mechanisms, ViTs are evaluated for their capability to resist perturbations and model intricate dependencies within medical images.
Adversarial Robust Pretraining: By pretraining models on adversarially augmented data, we aim to build resilient feature representations that withstand adversarial manipulations.

II. Motive

This project seeks to ensure the development of safe, robust, and reliable machine learning models tailored for critical healthcare applications. By strengthening models against adversarial attacks, we aim to:

Enhance trust: Build confidence in AI-driven diagnostics by ensuring consistent and accurate predictions.
Prevent misdiagnoses: Mitigate risks associated with adversarial vulnerabilities that could compromise patient

safety. Promote ethical AI: Maintain high standards of reliability and fairness in healthcare AI systems. The outcomes of this project not only contribute to building resilient AI systems for healthcare but also provide actionable insights into addressing adversarial threats in medical image classification. Robust and reliable AI models will pave the way for their widespread adoption, transforming the landscape of healthcare diagnostics and treatment planning.

III. Background

Adversarial robustness has emerged as a pivotal focus area in machine learning, particularly in high-stakes domains like healthcare. Neural networks, despite their remarkable performance in tasks such as image classification, are inherently vulnerable to adversarial attacks—carefully crafted perturbations that can mislead models into making incorrect predictions. This susceptibility raises serious concerns, especially in the healthcare sector, where erroneous predictions could lead to misdiagnoses and compromise patient safety. Adversarial attacks such as Projected Gradient Descent (PGD), Patch Attacks,Universal Adversarial Perturbations (UAPs), Evasion attacks, Model extraction attacks and Poisoning attacks are commonly used to evaluate and expose model vulnerabilities. While these attacks highlight critical weaknesses, they also serve as tools for enhancing model robustness. Training models to resist such perturbations forces them to learn more generalized and robust feature representations, ultimately making them more reliable. The healthcare sector demands the utmost accuracy and reliability from AI systems. Despite advancements in AI, many clinicians remain hesitant to adopt machine learning due to the potential for errors under adversarial conditions. A misdiagnosis, even in rare cases, can erode trust in AI systems and deter widespread adoption.

Adversarial robustness addresses this trust gap by ensuring that models:

- Maintain consistent performance even in the face of malicious perturbations.
- Reduce the likelihood of catastrophic errors in critical medical applications.
- Enhance interpretability and trustworthiness, fostering confidence among healthcare professionals.

Robust and safe machine learning models are not merely desirable but essential in building ethical and reliable AI systems for healthcare. Strengthening defenses against adversarial attacks ensures that AI tools can operate as reliable assistants, augmenting human decision-making and paving the way for broader integration into clinical workflows. This project contributes to this goal by evaluating state-of-the-art attacks and defenses, pushing the boundaries of robustness in medical image classification.

IV. Dataset

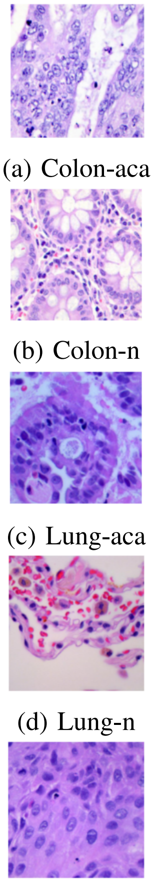
The dataset used in this project is derived from histopathological images of lung and colon tissues, specifically curated to represent five distinct classes:

- Lung Adenocarcinoma (Lung ACA): Malignant epithelial cells originating in glandular tissues of the lung.
- Lung Squamous Cell Carcinoma (Lung SCC): Cancer arising from squamous epithelial cells in the lung.
- Benign Lung Tissue (Lung N): Non-cancerous lung tissue.
- Colon Adenocarcinoma (Colon ACA): Cancer originating from glandular cells in the colon.
- Benign Colon Tissue (Colon N): Non-cancerous colon tissue.

The dataset was introduced in the paper *"A Large-Scale Dataset for Histopathological Image Classification, Segmentation, and Domain Adaptation"* [1]. It comprises high-resolution, expertly labeled histopathological images designed for classification tasks.

The dataset used in this project comprises over 25,000 high-resolution histopathological images of lung and colon tissues, which we have resized to 224x224 pixels during preprocessing to align with modern neural network architectures.

This dataset retains critical diagnostic details, such as cell morphology and tissue structure, enabling the differentiation of benign and malignant samples. Its real-world relevance lies in reflecting clinical challenges, such as subtle inter-class variations and high intra-class similarities, which mirror diagnostic complexities faced by pathologists. Automated classification on this dataset can enhance early detection and reduce the diagnostic burden on clinicians. Pre-processing steps include normalization, augmentation (e.g., rotation and flipping), and an 80-10-10 train-validation-test split to ensure robust evaluation. However, challenges such as adversarial vulnerabilities and the risk of misclassification due to inter-class similarities emphasize the importance of designing robust models.



(a) Colon-aca

(b) Colon-n

(c) Lung-aca

(d) Lung-n

(e) Lung-scc

Fig. 1: The five Distinct classes of the dataset

V. Methodology

In this project, we utilized a basic Convolutional Neural Network (CNN) for medical image classification on the Lung and Colon Cancer Histopathological Image Dataset. The CNN model was evaluated against three types of adversarial attacks: Projected Gradient Descent (PGD), Patch Attack, and Universal Adversarial Perturbation (UAP), each targeting different vulnerabilities. To defend against these attacks, we implemented adversarial fine-tuning, retraining the CNN with a combination of original and adversarially perturbed images. Additionally, Vision Transformers (ViTs) were employed to improve robustness, leveraging their self-attention mechanisms to capture global dependencies and resist localized perturbations. These methodologies aimed to evaluate and enhance the model's resilience to adversarial attacks, ensuring reliability in medical image classification tasks.

A. Convolutional Neural Network (CNN) for Medical Image Classification

1) Overview: A Convolutional Neural Network (CNN) was developed to classify histopathological images from the Lung and Colon Cancer Image Dataset. The CNN model was designed with simplicity and effectiveness in mind, consisting of multiple convolutional layers interspersed with max-pooling layers for spatial downsampling, followed by fully connected layers for classification.

The architecture of the CNN included the following layers:

- **Input Preprocessing**: Images were resized and normalized using a custom resize_scaling layer to standardize input dimensions and pixel value ranges.
- **Convolutional Layers**: The network used three sets of 2D convolutional layers with ReLU activation functions and kernel sizes of 3x3, followed by max-pooling layers to reduce dimensionality while retaining important features.
- **Batch Normalization**: A batch normalization layer was included to improve convergence and model generalization by normalizing activations during training.
- **Flattening and Dense Layers**: The feature maps were flattened and passed through two dense layers, with the final dense layer outputting predictions for the five classes using a softmax activation function.

The model was trained using the Adam optimizer with a learning rate scheduler and regularization techniques to ensure efficient training and prevent overfitting. The loss function used was sparse categorical cross-entropy, and accuracy was monitored as the primary performance metric.

To enhance training stability and performance, the following callbacks were used:

- **TensorBoard Logging**: For real-time monitoring of training metrics.
- **Model Checkpointing**: To save the best-performing model based on validation loss.
- **ReduceLROnPlateau**: To dynamically reduce the learning rate when the validation loss plateaued.
- **Early Stopping**: To terminate training early if the validation loss failed to improve for a set number of epochs, restoring the best weights.

This CNN served as the baseline model for evaluating the classification performance on the original dataset, as well as its robustness against adversarial attacks.

B. Projected Gradient Descent (PGD) Attack

1) Overview: Projected Gradient Descent (PGD) is one of the most widely used and powerful adversarial attack methods. It generates adversarial examples by iteratively perturbing the input images in the direction of the loss gradient, aiming to mislead the model into making incorrect predictions. The attack applies constraints on the perturbation magnitude to ensure it stays within a defined epsilon-ball, maintaining the perturbation's imperceptibility.

The PGD attack in this project was implemented with the following steps:

- **Initialization**: Start with a copy of the original image as the initial adversarial example.
- **Gradient Computation**: For each iteration, compute the gradient of the model's loss with respect to the current adversarial example.
- **Perturbation Update**: Modify the adversarial example by adding a small step (α) in the direction of the gradient's sign. This step ensures maximization of the model's loss for the true class.
- **Projection**: Clip the perturbation to ensure it stays within the allowed epsilon-ball, ensuring the adversarial example remains close to the original image.
- **Clipping to Image Range**: After each update, ensure pixel values of the adversarial example stay within valid image pixel ranges [0, 1].

The process repeats for a predefined number of iterations (*num_iter*) or until the adversarial image successfully causes misclassification.

2) Parameters:

- **Epsilon** (ε): Controls the maximum allowed perturbation magnitude. In this project, ε was set to 0.03.
- **Alpha** (α): Defines the step size for each perturbation update, set to 0.01.
- **Number of Iterations**: The attack ran for 5 iterations in the experiments, ensuring efficient generation of adversarial examples.

$$x_{t+1} = \text{clip}_{x,\epsilon}\big(x_t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(f_\theta(x_t), y))\big) \qquad (1)$$

where:

- x_t: Adversarial example at iteration t
- α: Step size
- ϵ: Perturbation bound
- f_θ: Neural network with parameters θ
- y: Ground truth label
- $\text{clip}_{x,\epsilon}$: Clips the adversarial example within the ϵ-ball of the original image x

C. Patch Attack

1) Overview: Patch attacks are a class of adversarial attacks where a visually perceivable and localized perturbation, often in the form of a small patch, is applied to input images. Unlike other adversarial attacks, patch attacks do not alter the entire image but focus on specific regions, making them effective against real-world systems where only parts of the input can be manipulated.

The patch attack methodology used in this project involves the following steps:

- **Patch Generation**: A small trainable patch of dimensions (10, 10, 3) was initialized with random values. The patch was optimized using the gradients of the model's loss to maximize misclassification probabilities.
- **Patch Application**: The patch was applied to images at a fixed position, blending it with the original image using a blending factor (*blend_alpha* = 0.05). The blending ensures that the patch seamlessly integrates into the image without introducing abrupt pixel transitions.
- **Optimization**: The patch was iteratively updated for 20 epochs using the Adam optimizer. Gradients were computed with respect to the patch to maximize the model's categorical cross-entropy loss for the true class.
- **Projection and Smoothing**: After each update, the patch values were clipped to the range [0, 1] to ensure valid pixel values. A Gaussian smoothing operation was applied to reduce noise and enhance the patch's effectiveness.
- **Evaluation**: The patch was applied to a test dataset, and the attack success rate was calculated by measuring the proportion of misclassified images.

2) Parameters:

- **Patch Size**: A small (10, 10, 3) patch was used, covering a localized region of the image.
- **Blending Factor**: A blending factor of 0.05 ensured smooth integration of the patch with the original image.
- **Learning Rate**: The patch optimization used a learning rate of 0.01.
- **Optimization Epochs**: The patch was optimized over 20 epochs.

$$x_{\text{patched}} = (1 - M \cdot \beta) \cdot x + M \cdot \beta \cdot P \qquad (2)$$

where:

- M: Binary mask defining the patch's location
- β: Blending coefficient
- x: Original image
- P: Adversarial patch

D. Universal Adversarial Perturbation (UAP)

1) Overview: Universal Adversarial Perturbations (UAPs) represent a powerful adversarial attack method where a single perturbation, applicable across all inputs, forces the model to misclassify diverse input samples. Unlike other attacks tailored for specific images, UAPs are model-agnostic and exploit the inherent vulnerabilities of the decision boundaries.

The following steps outline the generation and evaluation of UAPs:

- **UAP Initialization**: A universal perturbation tensor of size (*IMG_SIZE*, *IMG_SIZE*, 3) is initialized with zeros. This tensor is trainable and is iteratively updated during the attack.
- **Optimization Process**: The UAP is optimized to maximize the misclassification rate over the entire dataset:
  1) *Perturbation Application*: For each batch, the UAP is added to the original images while ensuring the perturbed pixel values remain within a valid range [0, 1].
  2) *Loss Computation*: The categorical cross-entropy loss is computed between the true labels and the model's predictions for the perturbed images. The loss is maximized to mislead the model.
  3) *Gradient Update*: Gradients of the loss with respect to the UAP are computed and used to update the perturbation using the Adam optimizer.
  4) After each update, the UAP is projected back to the ϵ-ball constraint ($\|\text{uap}\| \leq \epsilon$), ensuring that the perturbation remains within a predefined bound.
- **Evaluation**: The trained UAP is applied to the test dataset, and the accuracy of the model on the perturbed images is measured. This evaluates the model's vulnerability to UAPs and quantifies the attack's success.

2) Parameters:

- **Epsilon**: Maximum perturbation magnitude, ensuring the perturbation remains imperceptible to humans ($\epsilon = 0.05$).
- **Max Iterations**: The number of optimization iterations (*max_iters* = 10).
- **Learning Rate**: The step size for updating the UAP (*learning_rate* = 0.01).

$$\arg\min_{v} \mathbb{E}_{(x,y)\sim D}\big[\mathcal{L}(f_\theta(x+v), y)\big] \quad \text{subject to} \quad \|v\|_\infty \leq \epsilon \tag{3}$$

where:

- D: Data distribution
- v: Universal perturbation
- $\|v\|_\infty$: Perturbation norm bounded by ϵ
- f_θ: Neural network with parameters θ
- $\mathcal{L}$: Loss function
- **Adversarial Success**: UAPs achieved a high success rate, causing significant misclassifications across a diverse set of test images.
- **Visualization**: The generated UAP was visualized as a heatmap, showcasing patterns learned to exploit the model's vulnerabilities.

3) Implications: The success of UAPs demonstrates the existence of universal decision boundary flaws that can be exploited by adversarial attacks. This highlights the need for robust defenses capable of addressing global model weaknesses rather than individual image-level vulnerabilities.

E. Fine-Tuning the Baseline Model on Combined Attacked and Original Dataset

1) Overview: To enhance the robustness of the baseline Convolutional Neural Network (CNN) model against adversarial perturbations, we performed adversarial fine-tuning. This process involved pretraining the model on a merged dataset comprising both clean (original) images and adversarially attacked images. The goal was to improve the model's resilience to adversarial attacks.

Dataset Preparation:

- Original Dataset: Images from the five classes (*lung_scc, lung_aca, lung_n, colon_aca, colon_n*) were used as the clean dataset.
- Attacked Dataset: Adversarially perturbed images were generated. These images were organized into the same class-based folder structure as the original dataset, ensuring consistency.
- Merged Dataset: The original and attacked datasets were combined into a single dataset. Each clean image was paired with its corresponding adversarially perturbed counterpart.

Data Augmentation:

- Clean and adversarial images were concatenated and shuffled to ensure uniform training.

- Images were preprocessed with normalization and resized to a uniform shape of 180×180 pixels.

Fine-Tuning the Baseline CNN:

- The baseline CNN model, pretrained on the clean dataset, was used as the starting point.
- The model was fine-tuned using the merged dataset to improve its ability to classify images accurately under adversarial perturbations.

Training Process:

- For each batch, clean and perturbed images were passed through the model together.
- The Sparse Categorical Crossentropy loss function was used to calculate the combined loss.
- The Adam optimizer with a learning rate of 1×10^{-4} facilitated efficient training.
- Metrics such as accuracy and loss were tracked to monitor performance across epochs.

2) Model Evaluation: The fine-tuned model was evaluated on the following datasets:

- Clean Test Dataset: To measure the model's ability to classify original images post-training.
- Adversarial Examples: To assess its robustness against adversarial attacks.

The results were compared with the baseline model's performance to quantify the improvement.

3) Implications: Fine-tuning the baseline model on a combined dataset of clean and adversarially attacked images highlights the effectiveness of adversarial training as a defense mechanism. By exposing the model to a broader range of perturbations during training, it becomes better equipped to handle real-world adversarial challenges, improving its robustness and reliability in critical applications such as medical image classification.

F. Vision Transformer (ViT) Defense Mechanism

1) Overview: Vision Transformers (ViTs) serve as an effective defense mechanism against adversarial attacks due to their ability to model long-range dependencies in image features. Unlike Convolutional Neural Networks (CNNs), which focus on localized patterns, ViTs leverage self-attention mechanisms to process global context, making them more robust against adversarial perturbations such as Projected Gradient Descent (PGD), Universal Adversarial Perturbations (UAPs), and Patch attacks.

ViT Model Architecture:

- **Patch Embedding**: Input images were divided into non-overlapping patches. Each patch was embedded into a higher-dimensional space using a convolutional layer.
- **Positional Encoding**: To retain spatial information, positional encodings were added to the patch embeddings.
- **Transformer Encoder**: The core of the model consisted of a multi-head self-attention mechanism with 12 Transformer layers.
- **Classification Head**: A dense layer classified the images into the five medical image categories: *lung_scc, lung_aca, lung_n, colon_aca, colon_n*.

Training and Fine-Tuning:

- The ViT model was trained using a merged dataset of clean and adversarially perturbed images.
- Sparse categorical cross-entropy was used as the loss function.
- Early stopping and learning rate reduction strategies were applied.
- The model was trained for 40 epochs with a batch size of 64.

$$z_0 = [x_{\text{cls}}; x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{\text{pos}} \tag{4}$$

where:

- x_p^i: i-th patch of the image
- E: Patch embedding matrix
- E_{pos}: Positional embedding
- x_{cls}: Classification token

The attention mechanism used in the transformer is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5}$$

where:

- Q, K, V: Query, Key, and Value matrices
- d_k: Dimensionality of the keys

Performance Evaluation:

- The model was evaluated on both the clean test dataset and adversarially perturbed datasets.
- Metrics such as loss, accuracy, confusion matrix, and classification reports were computed to assess performance.
- Predictions on random samples were visualized to highlight successful and failed classifications.

2) Visualizations:

- **Training History**: Training and validation accuracy/loss were plotted to monitor convergence.
- **Confusion Matrix**: Provided a detailed breakdown of the model's performance across all classes.
- **Prediction Visualization**: Random samples with true and predicted labels highlighted correct and incorrect classifications.

3) Advantages of ViT as a Defense:

- **Global Context Understanding**: The self-attention mechanism captures dependencies across the entire image, reducing the impact of localized adversarial perturbations.
- **Robustness to Perturbations**: ViTs are inherently robust to noise and perturbations due to their reliance on global attention rather than localized features.
- **Scalability**: The patch-based approach makes ViTs scalable to larger datasets and higher-resolution images.

4) Results:

- **Training and Validation Performance**: Loss and accuracy trends across epochs indicated successful convergence.
- **Defense Robustness**: The fine-tuned ViT model demonstrated improved accuracy on adversarial datasets compared to the baseline CNN model.
- **Error Analysis**: The confusion matrix and visualizations highlighted areas for further improvement, particularly for challenging classes.

By leveraging the capabilities of Vision Transformers, the robustness of the classification pipeline was significantly enhanced, showcasing their effectiveness as a defense mechanism in adversarially attacked scenarios.

VI. Results

This section presents the performance evaluation of our model under adversarial attacks and the effectiveness of the implemented defense mechanisms.

The baseline CNN model achieved an accuracy of 0.97 on the clean test dataset, highlighting its capability to classify lung and colon histopathological images effectively under normal conditions.

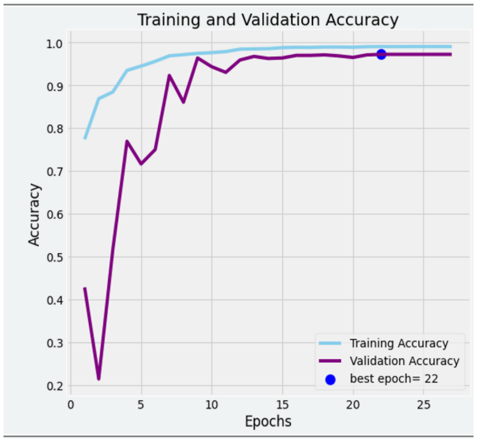


Fig. 2: Graph for Training and Validation Accuracy

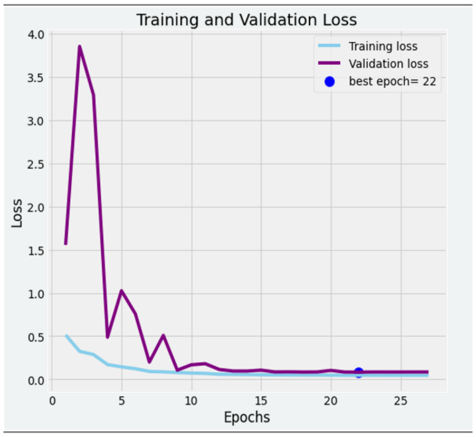


Fig. 3: Graph for Training and Validation Loss

The CNN model's accuracy dropped to 0.22 after implementing PGD attack, exposing its vulnerability to pixel-wise perturbations that exploit the loss gradient.

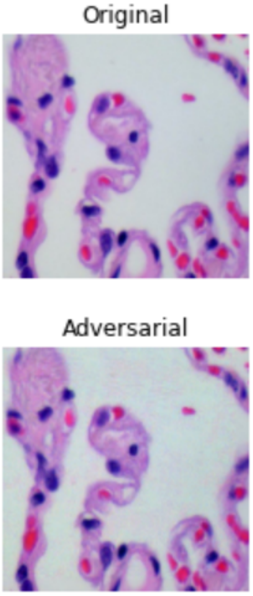


Fig. 4: PGD attack Results

The accuracy decreased to 0.26 after implementing Patch attack as shown in Fig-5, demonstrating the model's susceptibility to localized perturbations, simulating real-world adversarial scenarios.

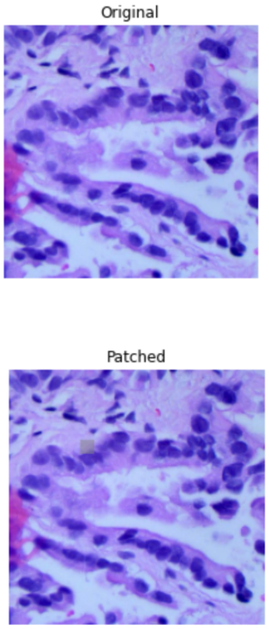


Fig. 5: Patch Attack Results

The model's accuracy was significantly impacted due to UAP attack, reducing to 0.21, indicating a lack of robustness to dataset-level perturbations.

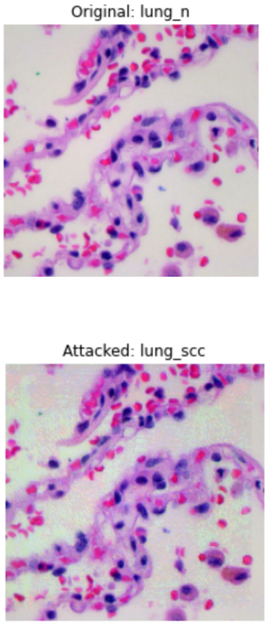


Fig. 6: UAP Attack Results

To establish a baseline for adversarial defense, we fine-tuned a robust model using the pretrained weights of our initial CNN model. This model was adapted to adversarially augmented datasets, including perturbations from PGD, Patch Attack, and UAP, balancing its performance on both clean and perturbed images with a weighted loss function. The fine-tuned model demonstrated improved resilience to adversarial attacks while retaining strong classification accuracy, serving as a benchmark for evaluating advanced defenses. Insights from this model informed our transition to Vision Transformers (ViT), leveraging their global self-attention mechanisms to address the limitations of CNNs in handling adversarial perturbations. By building on the robust baseline, we strategically refined defenses, ensuring scalable and effective protection against adversarial vulnerabilities in medical image classification.

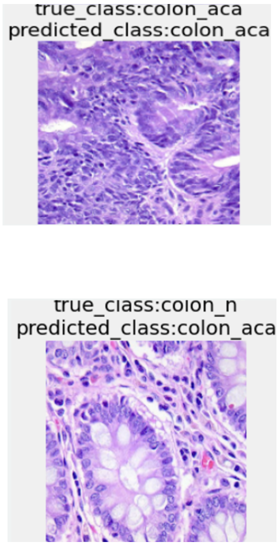
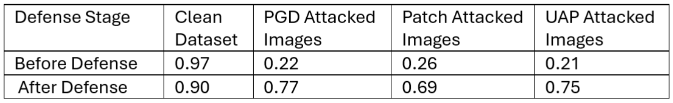


Fig. 7: Defense Mechanism Results



Defense Stage	Clean Dataset	PGD Attacked Images	Patch Attacked Images	UAP Attacked Images
Before Defense	0.97	0.22	0.26	0.21
After Defense	0.90	0.77	0.69	0.75

VII. Conclusion

In conclusion, this project highlights the critical importance of developing robust and reliable machine learning models for medical image classification, particularly in the context of lung and colon histopathology. By evaluating the vulnerabilities of convolutional neural networks (CNNs) to adversarial attacks, including PGD, Patch, and UAP, and implementing defenses through fine-tuned robust models and Vision Transformers (ViTs), the study underscores the potential for AI to enhance diagnostic accuracy in healthcare. The improved performance against adversarial attacks demonstrates the viability of integrating robust AI solutions into clinical settings, addressing the skepticism surrounding AI adoption in sensitive domains. This work provides a foundation for future research focused on enhancing model resilience and ensuring trustworthy AI systems for real-world healthcare applications.

VIII. Contributions

This project was a collaborative effort, with each team member bringing unique expertise to address the challenges of robust medical image classification under adversarial conditions.

Nikita Vinod Mandal focused on building and optimizing the initial CNN model, which served as the foundation for the project's baseline evaluations. Additionally, Nikita implemented and analyzed adversarial attacks, particularly the Patch Attack, providing critical insights into the model's vulnerabilities and areas for improvement.

Bhuvan Karthik Channagiri led the development of robust defense mechanisms, including the integration and fine-tuning a custom version of Vision Transformers (ViTs) as a defense strategy. Bhuvan also worked on establishing a robust baseline using pretrained CNN weights and iteratively enhancing the architecture to counter adversarial attacks effectively.

Together, they worked on evaluating the model's performance under PGD, Patch, and UAP attacks, ensuring a comprehensive approach to improving robustness and reliability for real-world medical AI applications.

IX. Future Work

For future work, we plan to explore Differentiable Architecture Search (DARTS) to automatically identify the optimal architecture for medical image classification tasks, particularly under adversarial conditions. DARTS is a gradient-based Neural Architecture Search (NAS) technique that efficiently discovers task-specific architectures by optimizing the search process using gradient descent. By incorporating adversarial accuracy into the search objective, this approach can identify architectures inherently robust against a variety of attack types. This eliminates the need for extensive manual design, allowing for a more systematic exploration of architectural possibilities. Once the optimal architecture is identified, we aim to fine-tune it further to enhance its robustness and performance across clean and adversarially perturbed datasets, paving the way for safer and more reliable AI applications in healthcare.

[1]

References

[1] "Lung and Colon Cancer Histopathological Images Dataset." [Online]. Available: https://arxiv.org/abs/1912.12142

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1706.06083

[3] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial Patch," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) Workshop*, 2017. [Online]. Available: https://arxiv.org/abs/1712.09665

[4] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal Adversarial Perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [Online]. Available: https://arxiv.org/abs/1610.08401

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[7] C. Xie, M. Tan, B. Gong, J. Wang, A. L. Yuille, and Q. V. Le, "Adversarial Examples Improve Image Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [Online]. Available: https://arxiv.org/abs/1911.09665

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. [Online]. Available: https://arxiv.org/abs/2005.12872

[9] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2017. [Online]. Available: https://arxiv.org/abs/1608.04644

[10] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble Adversarial Training: Attacks and Defenses," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1705.07204

[11] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan, "Theoretically Principled Trade-off between Robustness and Accuracy," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. [Online]. Available: https://arxiv.org/abs/1901.08573

[1] Project GitHub Link: https://github.com/NikkiMandal/Adversarial-attacks-and-defence-on-Medical-Data