Project Report
Classification and Robustness of Deep Learning Networks with Medical Images

Jon Van Veen
for Prof. Kassem Fawaz
ECE 697
University of Wisconsin-Madison

Abstract

Medical imaging deep learning models show enormous promise to improve medical diagnoses and make clinical workflows more efficient. However, dealing with the vulnerability of DL models to attacks from adversarial data is a primary concern to avoid attempts at insurance fraud or a false medical reimbursement claim. This project investigates classification performance of a vision transformer known as the Swin Transformer, with ResNet50 as a performance benchmark, on two large medical imaging datasets, and analyzes Swin-T's performance to noisy images as a first step towards an analysis of adversarial robustness.

1 Introduction

Transformers are neural networks that aim to learn meaning and context by analyzing relationships in sequential data. Transformers are the leading state-of-the-art method in the field of natural language processing (NLP) [1]. Vision transformers (ViTs) apply attention- and context-based learning to several common computer vision and image processing tasks, such as classification and segmentation, and have shown outstanding success at these tasks [1].

Classification and segmentation are of great interest in crucial medical imaging applications, such as tumor detection and radiation treatment therapy [2]. While much research exists analyzing ViT performance on traditional (RGB) images, there is much less investigation on ViT performance for medical images [2]. Obstacles to successful classification and segmentation in medical images include noise and unwanted artifacts, which often decrease model performance [2]. Additionally, an attacker may try to manipulate medical data with perturbations to profit from insurance fraud or a false medical reimbursement claim. It is therefore is critical that models used in clinical settings are robust against adversarial manipulation [2]. Prior research has demonstrated the vulnerability of deep learning models to such attacks [3]. While some prior literature evaluates robustness of transformers to traditional (RGB) images, as of 2022 no principled approach to evaluating transformer robustness in medical imaging modalities exists [2]. These reasons motivate the use of medical imaging data and the investigation of ViT robustness in this project. The intended contribution of this project relative to referenced related work is a comparison of ViTs to convolutional neural networks on medical imaging, as well as first steps towards the evaluation of adversarial robustness on ViTs.

ViTs have taken over as the state-of-the-art class of models from the more well-established convolutional neural networks (CNNs) for image processing tasks [1]. For several years in the 2010s CNNs led the field for these tasks, but in the last three to four years ViTs have handily outperformed CNNs in classification and segmentation tasks on common benchmark datasets such as Imagenet [1]. Because of their relatively recent rise to prominence in computer vision literature and state-of-the-art

performance, ViTs will likely continue to provide innovations in computer vision in the next few years [1][2].

This project references several recent related works. Khan et. al. provide a survey of recent ViT literature [1]. Shamshad et. al. provide a survey of ViTs applied to medical imaging tasks and point out current shortcomings and future directions for research [2]. Finlayson et. al. demonstrate the vulnerability of medical deep learning models to adversarial data [3]. Liu et. al. propose a popular ViT architecture known as the Swin Transformer that this project investigates [4]. He et. al. propose deep residual CNNs for imaging tasks, from which the ResNet50 model that this project investigates is derived [5]. Mao et. al. propose a robust ViT architecture for traditional RGB images [6]. Goodfellow et. al. introduce and motivate the problem of adversarial data manipulation [7]. Croce Hein propose AutoAttack, a benchmark ensemble of attacks to evaluate robustness [8]. Liu et. al. investigate robustness on a U-net architecture for medical image segmentation [9]. The two datasets used in this project are referenced at [10] and [11] respectively. Finally, the Adversarial Robustness Toolbox of many attacks and defenses was investigated [12].

2 Methods and Theory

Classification is the process of assigning labels to images as a whole, typically with only one object, while segmentation is the process of labeling specific regions of an image and determining where those regions are located [1]. ViTs typically approach these tasks by treating single patches in an image as the fundamental unit of analysis, similarly to how NLP transformers treat single words within a sentence as the fundamental unit of analysis. ViTs then try to determine the context of each patch with respect to the other patches and the image as a whole. This calculation is known as "attention" in transformer literature [1].

An insightful conceptual difference between ViTs and CNNs is as follows. Because CNNs work on the scale of very small image windows (e.g. 3x3), they pick out fine, high-frequency details in images, and act as highpass filters. In contrast, ViTs analyze over a global context, and so act as lowpass filters. Intuitively, this property of ViTs should make them more robust than CNNs to a set of common attacks that modify image data with high-frequency noise.

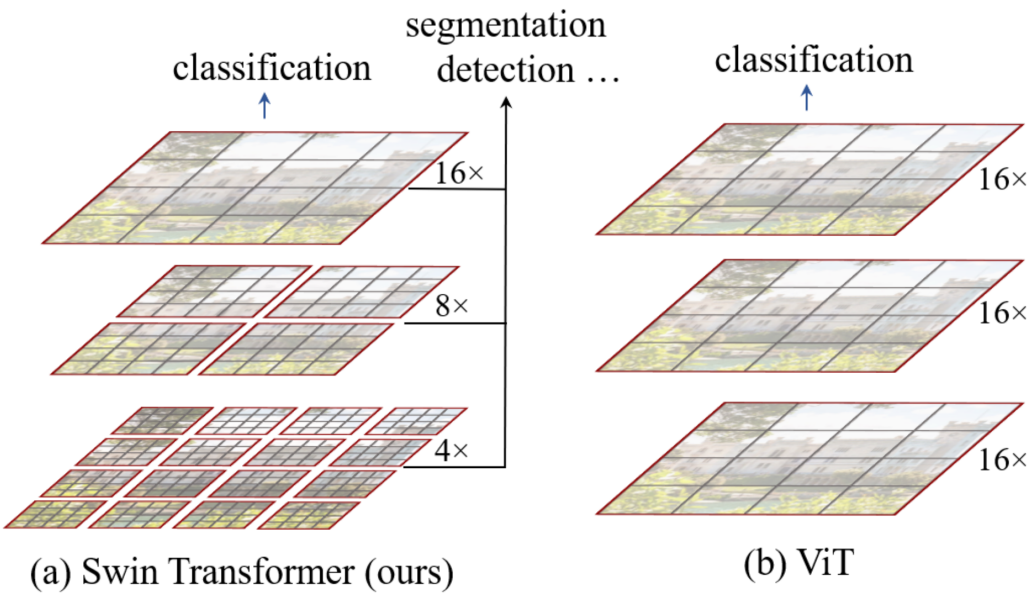

Figure 1: Visual comparison of standard ViT and Swin-T [3].

A popular recent ViT architecture is known as the Swin (shifted windows) Transformer [3]. Traditional ViTs analyze image patches by calculating attention across a global level at each layer in the network. In contrast, Swin-T uses hierarchical feature maps and shifted windows so that each layer calculates attention only within a certain grouping of image patches, outlined in red in fig. 1. Each grouping of patches is proportional to those in other layers. This scheme results in linear computational complexity, as opposed to quadratic complexity for standard ViTs, with comparable or better performance [3]. The improved computational complexity can be demonstrated by comparing the self-attention formulations for standard ViTs (1) and that of Swin-T (2) for an image containing h by w patches [3]:

$$\Omega = 4hwC^2 + 2(hw)^2C \tag{1}$$

$$\Omega = 4hwC^2 + 2M^2hwC \tag{2}$$

where C represents the projection of features to an arbitrary dimension, and each window over which self-attention is computed contains M by M windows with M fixed, e.g. 4 by 4 in figure 1. Note that (1) is quadratic with respect to a varying patch number hw, while (2) is linear because M is always fixed [3]. The advantages of Swin-T mean it is one of the leading models to use as a backbone for computer vision tasks [1].

Many different attacks to generate adversarial data exist [7][8]. Attacks apply small, visually imperceptible perturbations to input images that are intentionally designed to fool a model into a worst-case misclassification [7]. Goodfellow et. al. demonstrate an example "white-box" adversarial perturbation that maximizes the loss function of a model using that model's parameters and the sign of the cost function's gradient, known as FGSM [7]. AutoAttack is a benchmark framework that uses an ensemble of many parameter-free attacks that is computationally efficient and does not require re-tuning for each new defense evaluation [8]. This framework identified significant broken defenses in dozens of recently published machine learning and computer vision models [8]. Applying AutoAttack to leading ViT architectures trained on medical image data is therefore a compelling research direction.

2.1 Approach and Project Development

Throughout this project I received guidance from the course advisor, Prof. Kassem Fawaz. I am also grateful to Dr. Alan McMillan and Jinnian Zhang for providing project guidance and help with code implementation. Jinnian provided help with AUC implementation, adding noise to data for evaluation, modifying network architecture to deal with multi-class data, and plots in the Results section.

My intended approach to this project was to evaluate the classification performance of a few different ViT architectures on three medical imaging datasets, and also train a few CNN models on the same datasets as a comparison. Because ViTs are much less researched and have more potential for medical imaging than do CNNs, ViTs are the emphasis of the results in this report. After establishing the baseline non-adversarial performance, I intended to use AutoAttack to evaluate adversarial robustness for the two classes of models. Because published and well-used implementations already existed for the models I planned on investigating and the Autoattack framework, I assumed that putting together the data, models, and adversarial methods would be a plug-and-play process.

Howevever, that assumption did not pan out as the project progressed. The first problem I ran into was with the NIH Chest X-ray dataset. Individual images in this dataset often have multiple class labels corresponding to diagnoses of multiple diseases. However, the networks I planned on investigating were set up for single-label samples. This means that the network did not have enough information to accurately classify a validation image with multiple class labels into a single class. I initially ran the Swin-T without accounting for this problem, and obviously obtained nonsensical results. Jinnian helped me implement an additional sigmoid layer to the network, which employs a threshold activation value so that multiple classes can be predicted. This allowed the network to classify the validation images properly and resulted in reasonable baseline outputs.

Second, I found it very difficult to load my custom datasets into Pytorch in the way that my models and Autoattack were expecting. Online tutorials such as those on the official Pytorch website were of some help in this respect, but usually used toy examples of data that were neatly wrapped up in common Pytorch packages. I had to create custom Dataset and Dataloader objects from scratch in Pytorch, which was the source of many bugs and undesired results. I did eventually get this process to work and successfully ran my models using these custom functions. As the discussion so far indicates, I underestimated the amount of time and effort the much less exciting but always vital tasks of dataset organization and preparation would take.

Third, trying to put the Swin-T and the AutoAttack framework was much more challenging than I anticipated. Part of the reason is that no one as far as I was able to find had tried to implement adversarial methods on the Swin-T, so I found no examples that might have guided me in this process. The first AutoAttack implementation I attempted was that of the original author's Github page [8]. This involved wrapping two different "main" scripts together, which at times felt like trying to put a square peg into a round hole. After a lack of success here, I shifted to the AutoAttack implementation that is part of the Adversarial Robustness Toolbox [12]. Prof. Fawaz and two of his students were gracious enough to help me try to implement this approach. Unfortunately, I wasn't able to produce functioning robustness code with the time I had left in the semester.

A fourth, less significant obstacle was computing resources. My first experiments ran on the Euler cluster (detailed in the below section), which was busy with other jobs such that I often had to wait 24-48 hours for my own job to run, much less begin to debug or to start to see results. I did eventually have success running the Swin-T on the NIH dataset using Euler. I later shifted to running code using Google Colab Pro+. Despite its single GPU compared to Euler's eight GPUs, I found Colab trained and evaluated my models at least as fast or faster.

For the reasons listed above, I had to narrow my focus to just one representative model each between ViTs and CNNs. The models I chose were Swin-T and ResNet-50. I additionally narrowed my focus to two datasets instead of three.

At one point I had almost given up performing any kind of adversarial analysis. However, Alan and Jinnian suggested performing an evaluation by adding Gaussian noise to images during the validation process as an alternative. This analysis is an easy few lines of code added to the validation function. I acquired interesting results on the Swin-T architecture, which are shown and discussed in the following section. This is of course not a true, targeted adversarial robustness analysis, but I am still happy to have taken a first step in that direction in this project. Had I succeeded with the AutoAttack code, for each dataset I already sorted and planned on feeding 1000 images under an 80-20 train-test split to the AutoAttack classifier: 800 images to fit the classifier, and 200 images to predict and get robustness results. This subset of data seemed like a reasonable amount to ensure both meaningful and accurate robustness results as well as feasible computation time.

The root cause of my being unable to accomplish the original project objectives is likely my lack of any Pytorch experience prior to this project. There are many moving parts in Pytorch needed to put together a model and evaluate it in different ways, especially with a previously untried process of evaluating the Swin-T model on adversarial medical data using adversarial robustness packages. I believe I would have had more success in accomplishing the project goals if I had first had an organized introduction to Pytorch, such as through an online course. I imagine someone with significant Pytorch experience and intuition would have made much more progress than I did. Beyond Pytorch itself, previous experience coding with transformer architectures and adversarial robustness frameworks would also have been very beneficial.

2.2 Datasets and Implementation Details

Two medical imaging datasets were used to train and evaluate the Swin-T and ResNet50 models. The NIH Chest X-ray dataset contains 112,000 frontal-view chest x-ray images of the chest across 15

individual disease categories in .png format [10]. This dataset is very unbalanced, with the largest classes containing tens of thousands of samples and the smallest containing only one or two hundred. The nature of this dataset provides insight on how Swin-T performs on unbalanced datasets. As discussed above, this dataset has multi-class samples, so using it with networks intended to classify for single labels was a challenge. The second dataset I used is the ISIC skin cancer classification dataset, consisting of 71,000 .png images of benign and malignant skin lesions [11]. Because of the unbalanced nature of the NIH dataset, I trimmed the ISIC dataset to only use the top 10 classes in order to have a more balanced dataset for comparison. The resulting subset is still somewhat unbalanced, but much less so than the NIH dataset. Example images from both datasets are included in the below figures.

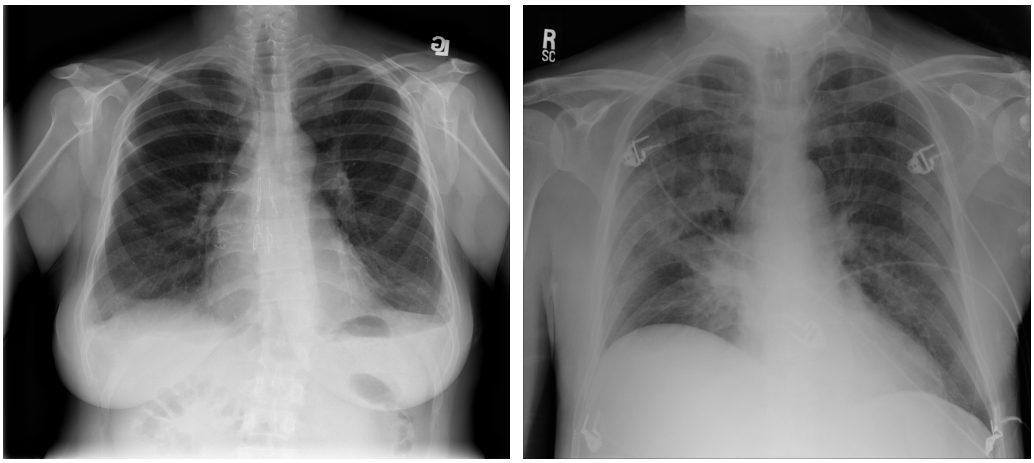

Figure 2: Two samples from the NIH Chest X-ray dataset [10].

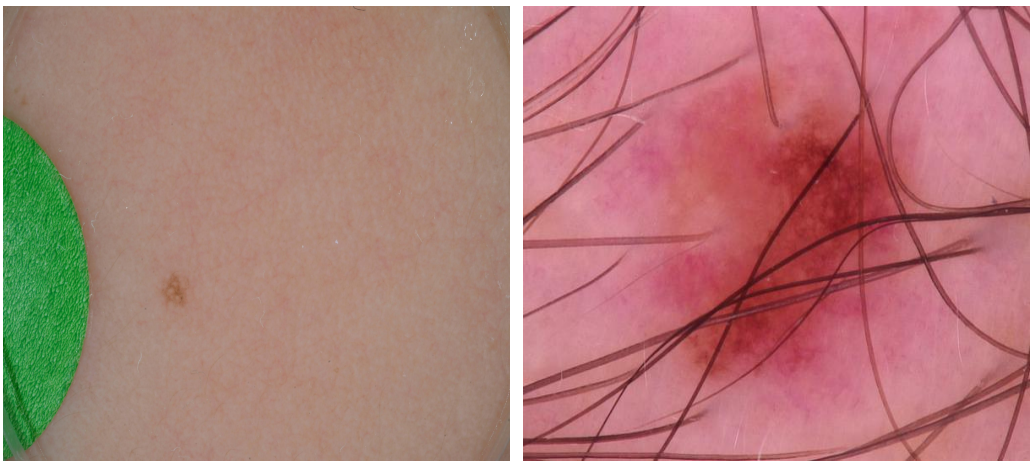

Figure 3: Two samples from the ISIC dataset [11].

Images from both datasets were downsampled to a resolution of 224 by 224, in keeping with common ViT research practice. Following the requirements of both models used, the datasets were organized into Imagenet's directory structure, consisting of train and validation folders with subfolders for each category. An 80-20 train-test split was applied to both datasets. Models were trained and evaluated using Colab Pro+, with a single GPU, as well as the Wisconsin Applied Computing Center's Euler cluster, with up to 8 GPUs. Models used pretrained Imagenet weights, and trained for 40 epochs on each dataset. All coding implementation used the Pytorch framework.

Noisy validation data was generated by adding Gaussian noise with zero mean to each validation image used in the clean classification task with Pytorch's randn() function, with variances chosen as described in the Results section. Checkpoints from the clean classification at 40 epochs were saved and then loaded before feeding the model noisy data in the validation loop.

3 Results

Performance metrics for the baseline classification task are shown in the below tables. These metrics include loss and accuracy on the validation dataset, accuracy on the top 1 and top 5 classes in each dataset, and AUC (Area Under the Receiver Operating Curve) scores.

NIH Chest X-ray Dataset

Model	Val Loss	Val Acc	Top 1 Acc	Top 5 Acc	AUC
Swin-T	**1.6676**	**48.4%**	48.4%	88.4%	0.8954
ResNet50	2.9967	41.7%	-	-	-

Figure 4: Swin-T and ResNet50 performance on the NIH Chest X-ray dataset.

ISIC Dataset

Model	Val Loss	Val Acc	Top 1 Acc	Top 5 Acc	AUC
Swin-T	**0.3074**	**89.1%**	89.1%	99.5%	0.9994
ResNet50	0.6236	86.4%	-	-	-

Figure 5: Swin-T and ResNet50 performance on the ISIC dataset.

Swin-T outperforms ResNet-50 in validation accuracy, and quite handily in validation loss. Because the superior performance of Swin-T was indicated by these metrics and prior literature, only these metrics were implemented for ResNet50. Swin-T shows lower scores on top-1 and top-5 accuracy for the NIH dataset than for the ISIC dataset, in keeping with expectations for a more unbalanced dataset with a greater number of classes. The AUC scores are very high; in the ISIC case, essentially 1. I note that the validation accuracy and top-1 accuracy scores are identical for these two results, as well as for the noisy evaluation that follows. The cause of this is unknown, but is probably not a significant problem given that the results are otherwise reasonable.

Following this evaluation, another evaluation was performed on noisy images from the validation set as described in the previous section. The range of values in validation image tensors was found to be 0.0 to 2.64 (compared to the standard 0 to 255 values in RGB images). A range of variances for the added noise was chosen accordingly to demonstrate the impact on performance of progressively more noise. The three variances thus chosen were 0.04, 0.10, and 0.40. Plots of the results are shown in figures 4 and 5 below, prepared with the help of Jinnian Zhang according to data shown in tables in the Appendix of this report. Additional data for the top-1 and top-5 accuracy metrics can be found in the Appendix tables.

3.1 Discussion

The clean data classification tables show that Swin-T exceeds the performance of ResNet50 for medical imaging data. This is in line with prior findings of the superior performance of ViTs

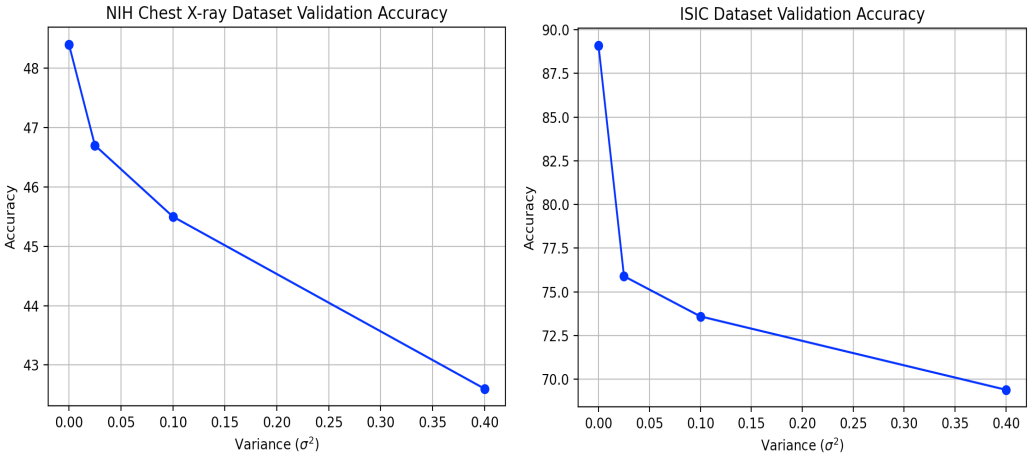

Figure 6: Swin-T validation accuracy vs. noise variance. (Left) NIH validation data. (Right) ISIC validation data.

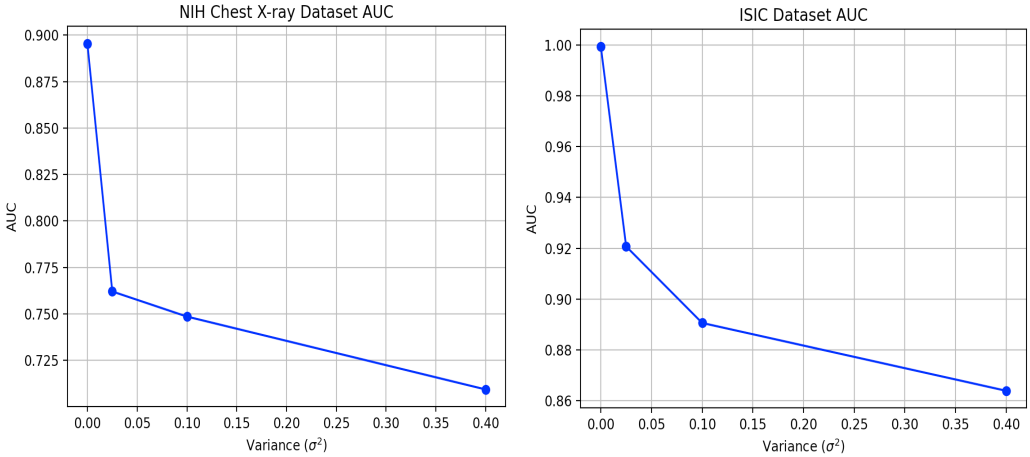

Figure 7: Swin-T AUC vs. noise variance. (Left) NIH validation data. (Right) ISIC validation data.

compared with CNNs on standard imaging data [1]. Using pretrained Imagenet weights with the models probably increased classification performance for both.

Results from evaluation on progressively noisier images shows the degradation in performance of Swin-T across all metrics investigated. The initial addition of noise caused a sharper decrease in validation accuracy on the ISIC dataset, with a smaller drop in accuracy with larger variances. On the NIH dataset, the validation accuracy dropped significantly less than that of the ISIC for the same variances. Conversely, the drop in AUC for the NIH data was much steeper than that for the ISIC data.

As Goodfellow et. al. note, random noise is a poor representative for true adversarial examples [7]. The applications for these results to true adversarial robustness are therefore probably quite limited. Nonetheless, these results empirically demonstrate the expected outcome of a very simple, non-targeted "attack" on a ViT trained in medical image domains.

4 Summary and Conclusions

My initial hypothesis beginning this project is that due to the lowpass-filter property of ViTs, they will be more robust than CNNs to common attacks. Unfortunately, I did not have sufficient experience and time to test this hypothesis in my project. However, the results I did produce using noisy validation data do not contradict the hypothesis. I expect that further research will follow a similar procedure to the one I followed in this project. First, baseline non-adversarial performance will need to be established for both the ViT and CNN classes of models on a medical imaging dataset. As demonstrated in this report, the performance of the ViT will exceed that of the CNN under this analysis. Next, adversarial training and evaluation should be performed using the same models and datasets, and the classification results compared to baseline results. With sufficient time and ViT/Pytorch experience, this line of research is very feasible, and I look forward to seeing this research in the near future.

Future research will consider adversarial robustness in medical image segmentation as well as classification. Segmentation is an inherently more difficult task than classification, and uses model architectures with notable differences compared to that of classification, such as the U-net used in [9]. Suitably large and diverse datasets curated specifically for segmentation will also be required, and acquiring such datasets is in itself a difficult task. An example of investigating adversarial robustness for segmentation can be found in [9].

Lastly, this project was an enlightening learning experience for me. I went into the project without any Pytorch experience or knowledge of ViTs or adversarial robustness methods. This semester I had the privilege of learning about and wrestling with these compelling frameworks and ideas. Even if I did not acquire the results I was hoping for, the experience and lessons learned were easily worthwhile.

5 References

[1] S. Khan, M. Naseer, M. Hayat, S. Zamir, F. Khan, M. Shah. *Transformers in Vision: A Survey.* ACM Computing Surveys, 2021. https://arxiv.org/abs/2101.01169

[2] F. Shamshad, S. Khan, S. Zamir, M. Khan, M. Hayat, F. Khan, H. Fu. *Transformers in Medical Imaging: A Survey.* arXiv, 2022. https://arxiv.org/abs/2201.09873

[3] S. Finlayson, H. Chung, I. Kohane, A. Beam. *Adversarial Attacks Against Medical Deep Learning Systems.* arXiv, 2018. https://arxiv.org/abs/1804.05296

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.* ICCV, 2021. https://arxiv.org/abs/2103.14030

[5] K. He, X. Zhang, S. Ren, J. Sun. *Deep Residual Learning for Image Recognition.* ILSVRC, 2015. https://arxiv.org/pdf/1512.03385.pdf

[6] X. Mao, G. Qi, Y. Chen, X. Li, R. Duan, S. Ye, Y. He, H. Xue. *Towards Robust Vision Transformer.* CVPR, 2022. https://arxiv.org/abs/2105.07926

[7] I. Goodfellow, J. Shlens, C. Szegedy. *Explaining and Harnessing Adversarial Examples.* ICLR, 2015. https://arxiv.org/pdf/1412.6572.pdf

[8] F. Croce, M. Hein. *Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks.* ICML, 2020. https://arxiv.org/abs/2003.01690

[9] Z. Liu, J. Zhang, V. Jog, P. Loh, A. McMillan. *Robustifying Deep Networks for Medical Image Segmentation.* J Digit Imaging, 2021. https://arxiv.org/abs/1908.00656

[10] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. Summers. *ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases.* CVPR, 2017. https://arxiv.org/abs/1705.02315

[11] International Skin Imaging Collaboration. *ISIC Archive.* ISIC, 2016. https://www.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery

[12] M. Nicolae, M. Sinn, M. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, B. Edwards. *Adversarial Robustness Toolbox v1.2.0.* CoRR, 2018. https://arxiv.org/pdf/1807.01069

6 Appendix

Below are tables for the Swin-T's performance on noisy validation data, from which the plots in the Results section are derived.

NIH Chest X-ray Dataset

Variance	Val Acc	Top 1 Acc	Top 5 Acc	AUC
No Noise	48.4%	48.4%	88.4%	0.8954
$\sigma^2 = 0.025$	46.7%	46.7%	86.8%	0.7621
$\sigma^2 = 0.1$	45.5%	45.5%	85.3%	0.7486
$\sigma^2 = 0.4$	42.6%	42.6%	81.0%	0.7094

Figure 8: Swin-T performance on progressively noisier NIH validation data.

ISIC Dataset

Variance	Val Acc	Top 1 Acc	Top 5 Acc	AUC
No Noise	89.1%	89.1%	99.5%	0.9994
$\sigma^2 = 0.025$	75.9%	75.9%	98.1%	0.9207
$\sigma^2 = 0.1$	73.6%	73.6%	98.6%	0.8906
$\sigma^2 = 0.4$	69.4%	69.4%	96.3%	0.8639

Figure 9: Swin-T performance on progressively noisier ISIC validation data.