Adversarial attacks and adversarial robustness in computational pathology

Narmin Ghaffari Laleh (1), Daniel Truhn (2), Gregory Patrick Veldhuizen (1), Tianyu Han (3), Marko van Treeck (1), Roman D. Buelow (4), Rupert Langer (5, 6), Bastian Dislich (5), Peter Boor (4), Volkmar Schulz (3, 7, 8), Jakob Nikolas Kather (1, 9, 10, 11)

(1) Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany
(2) Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany
(3) Department of Physics of Molecular Imaging Systems, Experimental Molecular Imaging, RWTH Aachen University, Aachen, Germany
(4) Institute of Pathology, University Hospital RWTH Aachen, Aachen, Germany
(5) Institute of Pathology, University of Bern, Switzerland.
(6) Institute of Pathology and Molecular Pathology, Kepler University Hospital, Johannes Kepler University Linz, Linz, Austria.
(7) Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany
(8) Comprehensive Diagnostic Center Aachen (CDCA), University Hospital Aachen, Aachen, Germany
(9) Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom
(10) Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany
(11) Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany

Abstract

Artificial Intelligence (AI) can support diagnostic workflows in oncology by aiding diagnosis and providing biomarkers. AI applications are therefore expected to evolve from academic prototypes to commercial products in the coming years. However, AI applications are vulnerable to adversarial attacks, such as malicious interference with test data aiming to cause misclassifications. Therefore, it is essential for the use of AI-based diagnostic devices to secure them against such attacks before widespread use. Unfortunately, no resistant systems exist in computational pathology so far.

To address this problem, we investigate the susceptibility of convolutional neural networks (CNNs) to multiple types of white- and black-box attacks. We demonstrate that both attacks can easily confuse CNNs in clinically relevant pathology tasks and impair classification performance. Classical adversarially robust training and dual batch normalization (DBN) are possible mitigation strategies but require precise knowledge of the type of attack used in the inference.

We demonstrate that vision transformers (ViTs) perform equally well compared to CNNs at baseline and are orders of magnitude more robust to different types of white-box and black-box attacks. At a mechanistic level, we show that this is associated with a more robust latent representation of clinically relevant categories in ViTs compared to CNNs.

Our results are in line with previous theoretical studies. We show that ViTs are robust learners in computational pathology. This implies that large-scale rollout of AI models in computational pathology should rely on ViTs rather than CNN-based classifiers to provide inherent protection against adversaries.

Introduction

Artificial intelligence (AI) with deep neural networks can extract clinically relevant information from digitized pathological slides of cancer.[1,2] Over the last several years, hundreds of studies have shown that diagnostic, prognostic, and predictive models can achieve accuracy which is comparable with gold standard methods.[3–6] Most studies investigate applications in cancer diagnostics and treatment, where pathological diagnosis is a cornerstone and slides are ubiquitous.[7–9] It is widely expected that AI systems will increasingly be used in clinical practice for cancer diagnostics and biomarker identification over the coming years.[10,11] Ultimately, such AI systems have the potential not only to make existing workflows more efficient, but also enable physicians to recommend improved treatment strategies for cancer patients.[12–14]

Considering this, it is crucial to ensure that the AI systems are robust before they are used in diagnostic routines. AI systems should be resilient to subtle changes in input data and yield a stable performance, even when the input signal is noisy. In particular, this includes *adversarial attacks* to the input signal, i.e. wilful modifications to the input data by a malicious actor. Adversarial attacks are a vulnerability of AI systems which is a concern in many domains.[15] The most common of these attack types are called *white-box attacks*. In such attacks, the adversary has full access to the model's parameters.[16] In contrast, *black-box attacks* hide the original model from the attacker. Adversarial changes to the original data are usually undetectable to the human eye but are disruptive enough to cause AI models to misclassify samples.

Cybersecurity is highly relevant for development and regulation of software in healthcare.[17] AI systems in healthcare are particularly vulnerable to adversarial attacks.[18] This poses a significant security risk: predictions of AI systems in healthcare have potentially major clinical implications, and misclassifications in clinical decision-support systems could have lethal consequences for patients. Thus, AI systems in healthcare should be particularly robust against any attacks. Yet, in computational pathology, only very few studies have explored adversarial attacks.[19] To date, no established strategy has been developed to make AI systems in the field of digital pathology robust against such attacks. The development of attack-resistant AI systems in pathology is therefore an urgent clinical need, which should ideally be resolved before these systems are widely deployed in diagnostic routine.

To date, convolutional neural networks (CNNs) are by far the most used type of deep neural network in digital pathology.[20] CNNs are capable of capturing high-level features such as edges from input data by applying various kernels throughout the training process. As of late 2020, vision transformers (ViTs) have emerged as an alternative to CNNs. ViTs use lower-dimensional linear embeddings of the flattened small patches extracted from the original image as an input to a transformer encoder.[21] Unlike CNNs, ViTs are not biased toward translation-invariance and locally restricted receptive fields.[22] Instead, their attention

mechanism allows them to learn distal as well as local relationships. Although ViTs have outperformed CNNs in some non-medical prediction tasks, the uptake of this technology is slow in medical imaging. To date, only very few studies have investigated the use of ViTs in computational pathology.[23,24] Technical studies have described improved robustness of ViTs to adversarial changes to the input data, but this has not been explored in medical applications.[25–27]

In this study, we investigated the robustness of CNNs in computational pathology toward different attacks and compared these results to the robustness of ViTs. Additionally, we trained robust models and evaluated their performances against the white- and black-box attacks. We analyzed the attack structure for both models and investigated the reasons behind their performances. We validated our results in two clinically relevant classification tasks in independent patient cohorts.

Materials and Methods

Ethics statement

This study was performed in accordance with the Declaration of Helsinki. We performed a retrospective analysis of anonymized patient samples. In addition to publicly available data from "The Cancer Genome Atlas" (TCGA, https://portal.gdc.cancer.gov), we used a renal cell carcinoma dataset by the University of Aachen, Germany (ethics board of Aachen University Hospital, No. EK315/19) and a gastric cancer dataset by the University of Bern (ethics board at the University of Bern, Switzerland, no. 200/14). This study adheres to the MI-CLAIM [28] checklist (**Suppl. Table 1**).

Patient cohorts

We collected digital whole slide images (WSI) of H&E-stained tissue slides of renal cell carcinoma (RCC) from two patient cohorts: TCGA-RCC (N=897 patients, **Suppl. Figure 1A**), which was used as a training set and AACHEN-RCC (N=249, **Suppl. Figure 1B**), which was used as a test set. The objective was to predict RCC subtypes: clear cell (ccRCC), chromophobe (chRCC), and papillary (papRCC). In addition, we obtained H&E-stained slides of gastric cancer from two patient cohorts: TCGA-GASTRIC (N=191 patients, **Suppl. Figure 1C**) for training, and BERN-GASTRIC (N=249 patients, **Suppl. Figure 1D**)[29] for testing. The objective was to predict the two major subtypes: intestinal and diffuse, according to the Laurén classification. Samples with mixed or indeterminate subtype were excluded. Ground truth labels were obtained from the original pathology report.

Image preprocessing

We tessellated the WSI into tiles (512 px edge length at 0.5 µm per pixel) which were color-normalized with the Macenko method.[30] No manual annotations were used. Background and blurry tiles were identified by having an average edge ratio smaller than 4, using the canny edge detection method, and were removed.[23] For each experiment, we selected 100 random tiles from each WSI. We used a classical weakly supervised prediction workflow[31,32] in which each tile inherited the ground truth label from the WSI and tile-level predictions were averaged over the WSI at inference. Before each training run, the total number of tiles per class was equalized by random downsampling.[2]

Experimental design

First, we trained deep learning models on categorical prediction tasks on the training cohort and validated the performance on the test cohort. We used two Deep Learning models, ResNet50[33], a convolutional neural network (CNN), and Vision transformers (ViT).[34] Then, we assessed the susceptibility of the trained

models towards white- and black-box adversarial attacks. Finally, we evaluated mitigation strategies against adversarial attacks. One strategy was to attack the images in the training cohort, termed adversarially robust training. The other strategy, specific to CNNs, was to use dual batch normalization, as introduced by Han et al. [35].

Implementation and analysis of adversarial attacks

For an image X belonging to class C_i, an adversarial attack perturbs X in such a way that the image is misclassified as $C_j, i \neq j$. We used four common types of attacks: (1) Fast Gradient Sign Method (FGSM)[36–38], a single-step gradient-based white-box attack; (2) Projected Gradient Descent (PGD)[39], a multi-step gradient-based white-box attack with attack strength ϵ; (3) Fast Adaptive boundary (FAB)[40], a more generic type of gradient-based white-box attack; and (4) Square attack[41], a black-box attack which places square-shaped updates at random positions on the input image. To measure which amount of noise is detectable by humans, we randomly selected three tiles from the AACHEN-RCC data set and attacked each of them with PGD with 50 attack strengths (0 to 0.5). We presented these tiles to a blinded human observer (medical doctor) who subjectively classified the images as “no noise detectable” and “noise detectable”. Subsequently, we determined the detection threshold by fitting a logistic regression model to the data. This analysis was run separately for noise generated with PGD on a ResNet and a ViT model. To visualize the adversarial noise, we subtracted the perturbed image from the original image, clipped at the $10th$ and $90th$ quantile for each color channel and scaled between 0 and 255. In addition, we visualized the latent space of deep layer activations of CNNs and ViTs. The activation feature vectors of ResNet50 (1×2048) and ViT (1×768) were reduced to (1×2) by Principal Component Analysis (PCA), and each component was scaled between 0 and 1. To quantify the separation between multiple classes in this latent space, we calculated the Euclidean distance[42] between all points of each class to the center of the corresponding classes and between the centers of classes.

Statistics

The main statistical endpoint was the patient-wise micro-averaged area under the receiver operating curve (AUROC). 95% confidence intervals were obtained by 1000-fold bootstrapping based on sampling with replacement. The test data set remained the same for the experiments between different models. All experiments were repeated five times with different random seeds. We reported the mean AUROC with standard deviation (SD) and median AUROC with interquartile range (IQR = $q_{75th} - q_{25th}$). Two-sided unpaired t-tests were used to compare sets of AUROCs between different deep learning models for the same experimental condition. No correction for multiple testing was applied. Furthermore, we calculated the attack success rate (ASR) using the predefined thresholds 0.3, 0.5, and 0.7. The ASR quantified the effectiveness of an attack by calculating the degree of misclassification: if the prediction score for the

perturbed image changed more than the threshold, the attack was deemed successful. The ASR was calculated for 50 randomly selected tiles per class from the AACHEN-RCC set.

Code availability

All source codes are publicly available: for image preprocessing[43], codes are available at https://github.com/KatherLab/preProcessing; for the baseline image analysis[23], codes are available at https://github.com/KatherLab/HIA and for adversarial attacks, codes are available at https://github.com/KatherLab/Pathology_Adversarial. Additional details are available in the **Supplementary Methods**.

Results

CNN and ViT perform equally well on clinically relevant classification tasks

Prediction of pathological subtypes of renal cell carcinoma (RCC) into clear cell (ccRCC), chromophobe (chRCC), and papillary (papRCC) is a widely studied task.[23,44] We trained ResNet, a convolutional neural network (CNN, **Figure 1A**) and a ViT (**Figure 1B**) on this task in TCGA-RCC (N=897 patients, **Suppl. Figure 1A**). The resulting classifiers performed well on the external test set AACHEN-RCC (N=249, **Suppl. Figure 1B**), reaching a mean area under the receiver operating curve (AUROC) of **0.960 [± 0.009]**. ViT reached a comparable AUROC of **0.958 [± 0.010]** (**Figure 1C, Suppl. Table 2**), which was not significantly different from the ResNet (p=0.98). The image tiles which were assigned the highest scores showed typical patterns for each subtype, demonstrating that ResNet and ViT can learn relevant patterns and generalize to an external validation cohort (**Figure 1D**). In addition, we evaluated the baseline performance of CNN and ViT on subtyping of gastric cancer.[32,45] When trained on the TCGA-GASTRIC cohort (N=191 patients, **Suppl. Figure 1C**) and tested on the BERN cohort (N=249 patients, **Suppl. Figure 1D**), CNN and ViT achieved mean AUROCs of **0.782 [± 0.014]** and **0.768 [±0.015]** respectively (**Figure 1E, Suppl. Table 2**). Again, the highest scoring tiles showed morphological patterns which are representative for the diffuse and intestinal subtype (**Figure 1F**).[46,47] Together, these data are in line with previous evidence[23] and show that CNNs and ViTs perform equally well for classification tasks in our experimental pipeline.

CNNs are susceptible to white-box adversarial attacks

We attacked CNNs with adversarial attacks (**Figure 2A**), evaluating white-box and black-box attacks (**Figure 2B**). By default, we used the most commonly used gradient-based attack Projected Gradient Descent (PGD), and additionally tested three other types of adversarial attacks (Fast Gradient Sign Method [FGSM], Fast Adaptive boundary [FAB], and Square attacks, **Figure 2C**). We found that with an increasing attack strength ε, the amount of visible noise on the images increased (**Figure 2D**). We quantified this in a blinded observer study and found that the detection threshold for adversarial attacks was ε=0.19 for ResNet models and ε=0.13 for ViT (**Suppl. Table 3, Suppl. Figure 2A-B**). With increasing attack strength, the classifier performance on the test set decreased. Specifically, we attacked with PGD with a low (ε = 0.25e-3), medium (ε = 0.75e-3) and high (ε = 1.50e-3) attack strength. The AUROC for RCC subtyping dropped from a baseline of **0.960** to **0.919, 0.749**, and **0.429** (**Figure 3A, Suppl. Table 4**). Similarly, when attacked with FGSM, the AUROC for RCC subtyping dropped from a baseline of **0.960** to **0.943**, **0.893**, and **0.782** (**Suppl. Table 5**). For the secondary classification task, subtyping of gastric cancer, the CNN models were even more susceptible to adversarial attacks. Here, the PGD completely degraded classification performance. The AUROC reached by the CNN dropped from a baseline of **0.782**

to **0.380**, **0.029** and **0.000** for the images attacked with low, medium and high ε (**Figure 3B, Suppl. Table 6**). Together, these data show that CNNs are highly susceptible to adversarial attacks in computational pathology

Adversarially robust training partially hardens CNNs

We subsequently investigated two possible mitigation strategies to rescue CNN performance. First, we evaluated adversarially robust training, in which PGD is applied to the training dataset so that the CNN can learn to ignore the noise patterns. Although training a CNN with PGD-attacked images (ε = 0.005) slightly reduced the RCC classification performance from at baseline from **0.960** to **0.953 (Suppl. Table 2)** it improved the model's robustness to attacks. For the PGD attack at inference, this adversarially robustly trained CNN yielded an average AUROC of **0.950**, **0.943,** and **0.931** for low, medium and high ε, respectively **(Figure 3A, Suppl. Table 7).** FGSM-attack at inference resulted in **0.952**, **0.949**, and **0.943 (Table1, Suppl. Table 8)** for RCC classification. Second, we investigated if the effect of adversarially robust training of CNNs could be enhanced by using a dedicated technique, dual-batch-normalized (DBN). The baseline performance of this model was an AUROC of **0.942 [± 0.022]** (p = 0.35) for RCC classification, which was not significantly inferior to the original model (**Suppl. Table 2**). For both attack types, DBN-CNN conveyed good protection at inference, but did not beat the normal adversarially robust training (**Figure 3A, Suppl. Table 7, Suppl. Table 8**).

In the secondary prediction task, adversarially robust training slightly lowered the classification accuracy at baseline (on non-attacked images) from **0.782 [± 0.014]** to **0.736 [± 0.01]**, but mitigated the vulnerability to attack, resulting in AUROCs of **0.724**, **0.698**, and **0.662** for low, medium and high ε (**Suppl. Table 9**). Together, these data show that the attackability of CNNs can be partly mitigated by adversarially robust training. Dual batch normalization (DBN) did not convey any additional robustness to CNNs.

ViTs are inherently robust adversarial attacks

We attacked ViTs with adversarial attacks and found that they were relatively robust against PGD and FGSM without any adversarial pretraining and without any modifications to the architecture. For low, medium and high PGD attack strengths in RCC classification, ViT AUROCs were slightly reduced from a baseline **0.958** to **0.944**, **0.908** and **0.827** (**Suppl. Table 4**), but ViT was significantly more robust than Resnet (p = **0.06**, **0.04**, and **0.01**). Similarly, for FGSM, AUROCs were slightly reduced from a baseline of **0.958** to **0.952**, **0.937**, and **0.926** (**Suppl. Table 5**). Also for FGSM, ViT was significantly more robust than ResNet (p = **0.23**, **0.04**, and **0.05**). For the secondary prediction task of gastric cancer subtyping, the baseline performance was lower for all classifiers when compared to RCC (**Figure 3B**). Also in this task, ViTs were significantly more robust to attacks than ResNet (p <= 0.01 for low, medium and high attack strength, **Suppl. Table 6**). Training a ViT in an adversarially robust way slightly reduced the baseline

performance for RCC classification from **0.958 [± 0.01]** to **0.937 [± 0.007]** (**Figure 3A**), and reduced the performance of ViT under a low-intensity PGD attack from **0.944 [± 0.011]** to **0.932 [± 0.008]**. However, for medium and high intensity attacks, adversarially robust training was beneficial for ViTs, slightly increasing the AUROC from **0.908 [± 0.015]** to **0.921 [± 0.01]** and from **0.827 [± 0.032]** to **0.903 [± 0.016]**, respectively (**Suppl. Table 4, Suppl. Table 7**). Similarly, in the gastric cancer classification task, adversarially robust training hardened ViTs: they only slightly reduced their baseline AUROC of 0.737 to 0.725, 0.698, and 0.657 under low, medium and high-intensity attacks, respectively (**Suppl. Table 9**). Next, we investigated whether the improved higher robustness of ViTs compared to CNNs extended beyond to another type of white-box attack, FAB, as well as a black-box attack, the Square attack. To this end, we selected 150 tiles from the RCC subtyping task and calculated the attack success rate (ASR) for nine experimental conditions: a liberal, a medium and a strict evaluation criterion, for a wide range of attack intensities with $\varepsilon = 0.005$, $\varepsilon = 0.01$ and $\varepsilon = 0.2$ (**Table 1**). For all four types of attacks, in baseline models and adversarially trained models, ViTs had the lower (better) ASR in the majority of experiments. For baseline models, ViT outperformed ResNet 9:0 (FGSM), 5:4 (PGD), 9:0 (Square) and 7:2 (FAB). For adversarially trained models, the margin was smaller, but ViT still outperformed ResNet in all experiments: 9:0 (FGSM), 5:4 (PGD), 9:0 (Square) and 7:2 (FAB) (**Table 1**).

To identify potential reasons for this higher robustness of ViTs towards adversarial attacks, we analyzed the adversarial noise obtained with white-box attacks on ViTs and ResNets (**Suppl. Figure 3A**). Quantitatively, we found that the magnitude of the gradients was consistently lower for ViT than for ResNet. Qualitatively, in ViT, there is a clear patch partition boundary alignment while ResNet patterns are more spatially incoherent (**Suppl. Figure 3B**). We hypothesize that this reflects the patch-based nature of ViTs, which causes learned features to contain less low-level information such as lines and edges from an input image and therefore making them less sensitive to high-frequency perturbations. In addition, we analyzed the structure of the latent space of the deep layer activations in ResNet and ViT. We found that for the original images in the RCC classification tasks, the instances in the classes were visually more clearly separated for ViT than for the CNN (**Figure 3C**). This was confirmed in the more difficult task of gastric cancer subtyping, in which also a clearer separation was seen (**Figure 3D**). Quantitatively, the instances within a given class were aggregated more tightly in the ViT latent space, and the distance between the centers of the classes were larger (**Suppl. Table 10**). When we attack the images with $\varepsilon = 0.05$ and used the baseline model to extract the features, the differences were even more pronounced: the ResNet latent space was more de-clustered than the ViT latent space (**Figure 3C-D**). We conclude that the high robustness of ViT towards white-box adversarial attacks when compared with CNN is associated with a better separation of distinct classes in the latent space.

Discussion

Medical software can be a target of cyberattacks, which have the potential to cause significant harm.[17] Adversarial attacks can manipulate AI systems into giving false predictions.[18] The number of AI systems used in healthcare is massively increasing.[48] A particularly relevant domain of application is computational pathology, where AI systems have been shown to solve clinically relevant questions in the last few years.[3] Based on these academic developments, AI products will likely enter the market in the near future. This process offers potential benefits in terms of efficiency and resource savings for diagnostic stakeholders, while at the same time offering the possibility of improved biomarkers for cancer patients. However, during this potential large-scale rollout of AI systems, it is important to ensure cybersecurity.[49]

Here, we show that CNNs in computational pathology are susceptible to adversarial attacks far below the human perception threshold. We show that existing mitigation strategies such as adversarial training and DBN do not provide universal mitigation. Addressing this issue, we explored the potential of ViTs to confer adversarial robustness to AI models. We show that ViTs perform on par with CNNs at baseline, and that they seem inherently more robust against adversarial attacks. Although no AI models are universally and fully attack-proof, our study demonstrates that ViTs seem much more robust against common white-box and black-box attack types, and that this is associated with a more robust behavior of the latent space compared to CNNs. Our findings add to a list of theoretical benefits of ViTs over CNNs and provide an argument to use ViTs as the core technology for AI products in computational pathology. Also, our findings are in line with studies in non-medical domains which analyzed robustness of ViTs in technical benchmark tasks.[50,51]

A limitation of our study is the restriction to cancer use cases and classification tasks. A more difficult task such as predicting the response to therapy would have even more severe clinical implications and could not even be directly checked by a pathologist (as could the diagnostic classification tasks used in the study), since negative consequences for prognostic misclassifications have a time delay. Future work should also address other types of adversarial attacks, such as physical-world attacks[15] or one-pixel attacks.[52] The uptake of newer AI models, such as text-image models, could also open vulnerabilities towards new types of adversarial attacks.[53] As multiple AI systems are nearing the diagnostic market, hardening these tools against established and emerging adversarial attacks should be a priority for the computational pathology research community in academia and industry.[18]

Figures

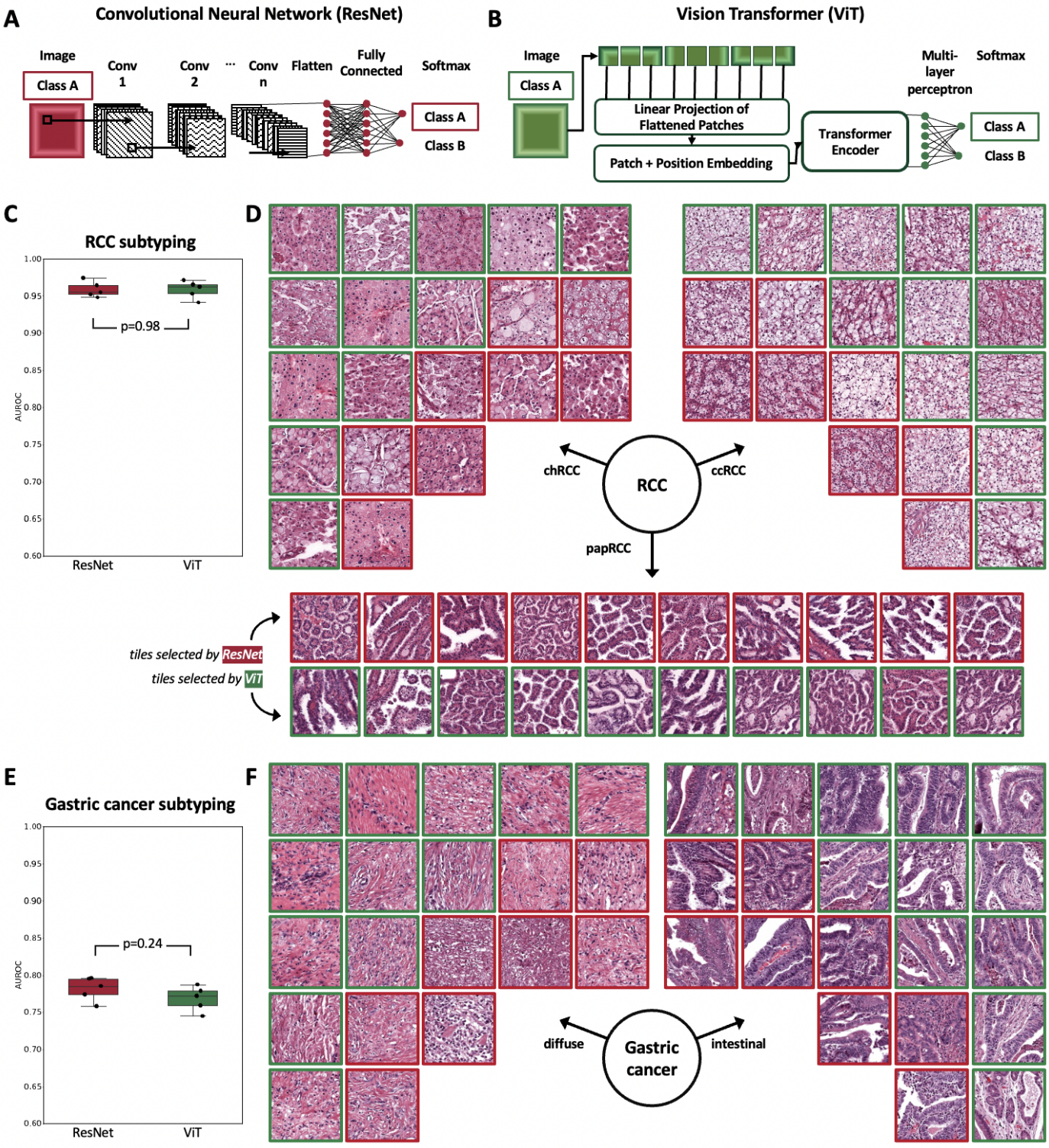

Figure 1: Cancer subtyping with Deep Learning. (A) Image classification with ResNet, **(B)** with a Vision Transformer (ViT). **(C)** Area under the receiver operating curve (AUROC) for subtyping of renal cell carcinoma (RCC) into clear cell (cc), chromophobe (ch) and papillary (pap). The box shows the median and quartiles of five repetitions (points) and the whiskers expand to the rest of the distribution. **(D)** Representative highly scoring image tiles for RCC, as selected by ResNet and ViT. **(E)** AUROC for subtyping of gastric cancer. **(F)** Highly scoring image tiles for gastric cancer, as selected by ResNet and ViT.

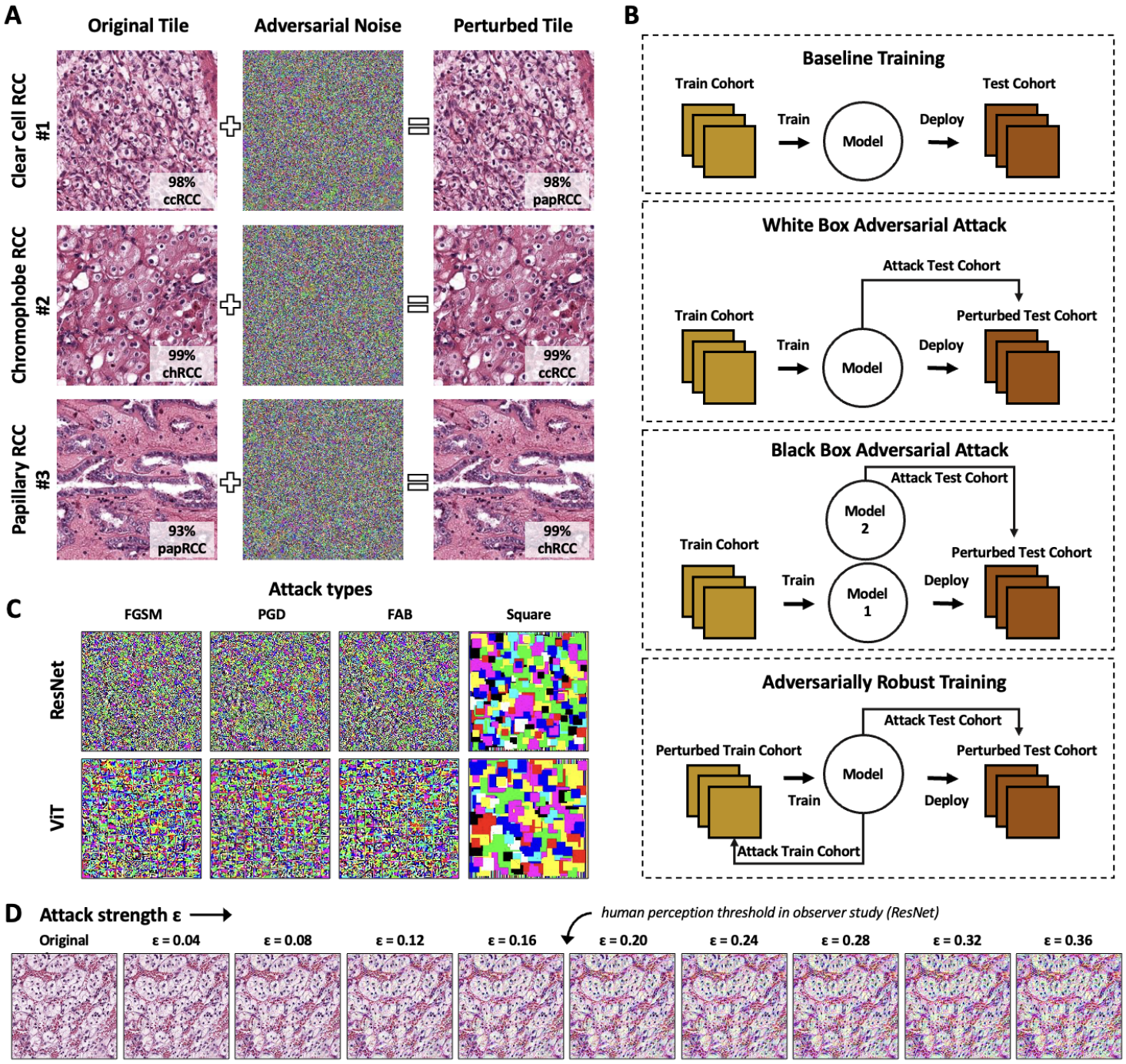

Figure 2: Adversarial attacks on computational pathology. (A) Adversarial attacks add noise to the image and flip the classification of renal cell carcinoma (RCC) subtyping into clear cell (cc), chromophobe (ch) and papillary (pap). The model's prediction confidence is shown on each image. **(B)** Experimental design for the baseline (normal) training, white-box and black-box attacks and for adversarially robust training. **(C)** Different attack algorithms yield different noise patterns. We used the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Fast Adaptive boundary (FAB) and Square attacks. **(D)** The attack strength ε increases the amount of noise which is added to the image. The average threshold for human perception is ε=0.19 for ResNet.

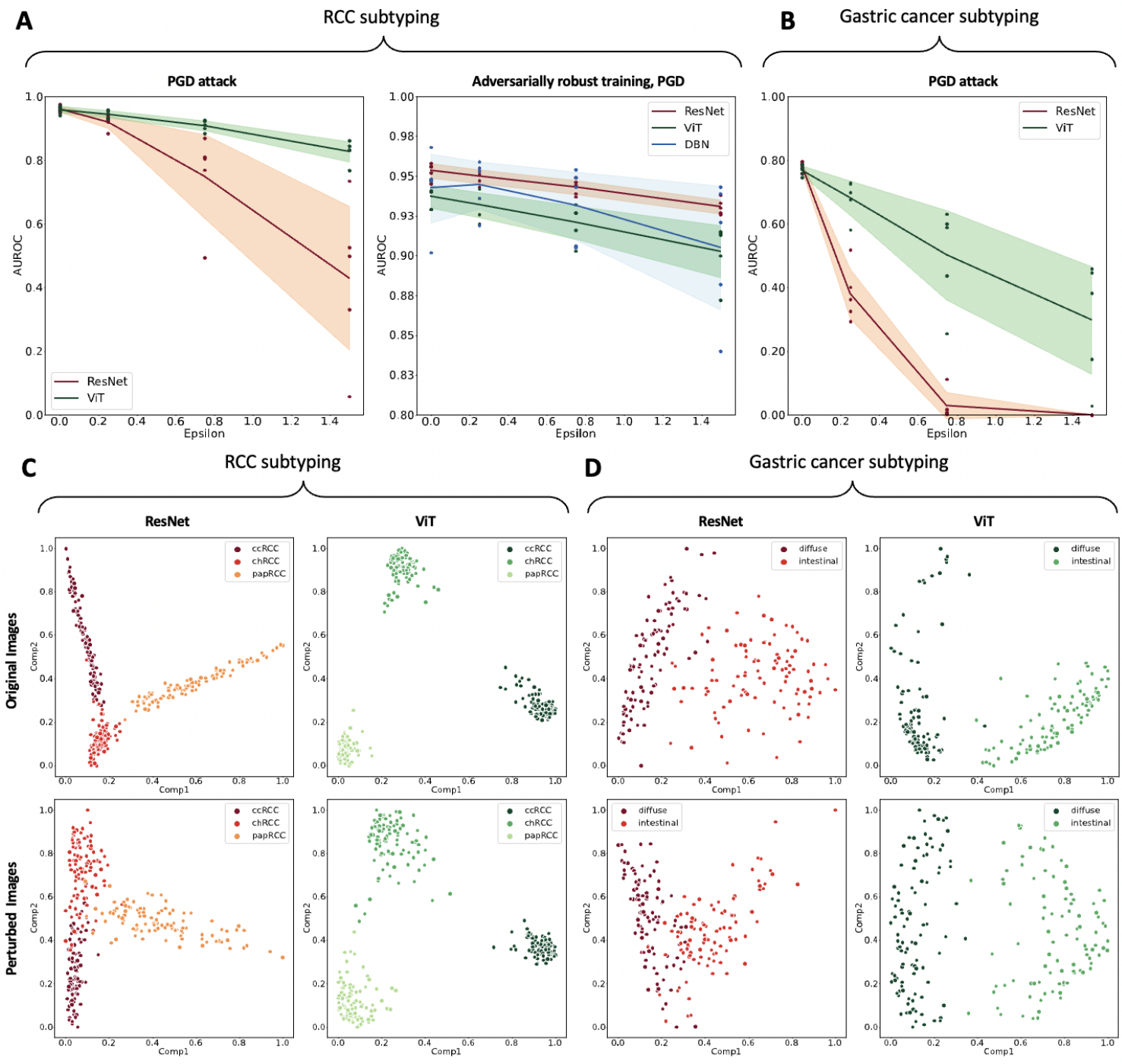

Figure 3: Vision Transformers are more robust to adversarial attacks than convolutional neural networks. **(A)** Micro-averaged AUROC for ResNet and ViT under PGD attack for RCC subtyping without (left) and with (right) adversarially robust training. Epsilon * 10E-3**. (B)** AUROC for ResNet and ViT for gastric cancer subtyping. ε * 10e-3. **(C)** First two principal components of the latent space of ResNet and ViT before (original) and after attack (perturbed) for RCC subtyping, for 150 highest-scoring image tiles. ViT has a better separation of the clusters before the attack and its latent space retains its structure better after the attack. **(D)** Latent space for the gastric cancer subtyping problem.

Tables

ε	T, t	Normal models								Adversarially trained models							
		FGSM		PGD		Square		FAB		FGSM		PGD		Square		FAB	
		ResNet	ViT	ResNet	ViT	ResNet	ViT	ResNet	ViT	ResNet	ViT	ResNet	ViT	ResNet	ViT	ResNet	ViT
0.005	T=0.3	73.3 %	**15.3 %**	77.3 %	**71.3 %**	32.0%	**0.00 %**	34.7 %	**16.7 %**	**0.00%**	**0.00%**	4.00 %	**2.00 %**	**0.00%**	**0.00%**	0.70 %	**0.00%**
	T=0.5	52.7 %	**0.70 %**	60.7 %	**33.3 %**	10.7 %	**0.00 %**	10.0 %	**2.00 %**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
	T=0.7	33.3 %	**0.00 %**	46.7 %	**10.7 %**	0.00 %	**0.00 %**	0.00 %	**0.00 %**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
	t [sec]	*7*	*10*	*42*	*45*	*1068*	*4154*	*796*	*772*	*11*	*14*	*31*	*48*	*3656*	*5100*	*704*	*707*
0.010	T=0.3	75.3%	**44.7 %**	**77.3 %**	83.3 %	40.0 %	**4.70%**	34.7 %	**28.0 %**	2.70 %	**2.00 %**	28.7 %	**22.0 %**	**0.00%**	**0.00%**	6.70 %	**3.30 %**
	T=0.5	54.7 %	**12.7 %**	**61.3 %**	64.0 %	18.0 %	**0.00 %**	10.7 %	**9.30 %**	**0.00%**	**0.00%**	9.30 %	**2.70 %**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
	T=0.7	42.7 %	**2.70 %**	47.3 %	**39.3 %**	1.30 %	**0.00 %**	0.00 %	**0.00 %**	**0.00%**	**0.00%**	0.70 %	**0.00 %**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
	t [sec]	*8*	*10*	*49*	*45*	*422*	*3359*	*796*	*766*	*7*	*9*	*41*	*45*	*600*	*4712*	*737*	*707*
0.200	T=0.3	74.7 %	**65.3 %**	**77.3 %**	84.0 %	54.0 %	**42.7%**	34.7 %	**30.7 %**	68.0 %	**56.0 %**	88.7 %	**86.0 %**	38.7 %	**32.7 %**	25.3 %	**23.3 %**
	T=0.5	56.0 %	**38.0 %**	**62.7 %**	69.3 %	50.0 %	**25.3 %**	**10.0 %**	12.0 %	40.7 %	**28.7 %**	78.7 %	**63.3 %**	23.3 %	**16.7 %**	**6.70 %**	**6.70 %**
	T=0.7	44.0 %	**17.3 %**	52.0 %	**47.3 %**	39.3 %	**1.30 %**	**0.00 %**	**0.00 %**	20.0 %	**6.00 %**	48.7 %	**34.7 %**	8.00 %	**1.30 %**	**0.00 %**	**0.00 %**
	t [sec]	*7*	*7*	*43*	*40*	*7*	*18*	*787*	*643*	*7*	*7*	*40*	*39*	*8*	*59*	*734*	*589*
N winners		0	9	4	5	0	9	2	7	5	9	2	9	6	9	6	9

Table 1: ViTs are more robust to adversarial attacks than ResNets, as measured by the Attack Success Rate (ASR), for the RCC classification task. The threshold T is the amount by which the model's predictions have to be shifted to consider an attack successful. T=0.3 is a liberal threshold and T=0.7 is a strict threshold. The computation time t is the time needed to apply the attack to 150 tiles. For pairwise comparisons between ResNet and ViT for the same experimental condition, the one with the lower (better) ASR is printed in bold. The bold digits are summed up for each column in the "N winners" row. In this experiment, 150 randomly selected tiles from AACHEN-RCC were used (same tiles for all experiments).

Additional information

Author contributions

NGL, DT and JNK designed the study; NGL and JNK developed the software; NGL performed the experiments; NGL, DT, TH, PB, and JNK analyzed the data; NGL, MVT performed statistical analyses; RDB, HB, TY, JCC, MH, PB provided clinical and histopathological data; all authors provided clinical expertise and contributed to the interpretation of the results. NGL, DT, GPV, and JNK wrote the manuscript and all authors corrected the manuscript and collectively made the decision to submit for publication

Conflicts of interest

JNK declares consulting services for Owkin, France and Panakeia, UK. No other potential conflicts of interest are reported by any of the authors.

Funding

JNK is supported by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and the Max-Eder-Programme of the German Cancer Aid (grant #70113864). PB is supported by the DFG, German Research Foundation (Project-IDs 322900939, 454024652, 432698239, 445703531, 445703531), European Research Council (ERC; Consolidator Grant AIM.imaging.CKD, No 101001791), Federal Ministry of Education and Research (STOP-FSGS-01GM1901A), and Federal Ministry of Economic Affairs and Energy (EMPAIA, No. 01MK2002A).

References

1. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
2. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
3. Echle, A. *et al.* Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br. J. Cancer* 1–11 (2020).
4. Schneider, L. *et al.* Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review. *Eur. J. Cancer* **160**, 80–91 (2022).
5. Kuntz, S. *et al.* Gastrointestinal cancer classification and prognostication from histology using deep learning: Systematic review. *Eur. J. Cancer* **155**, 200–215 (2021).
6. Nam, D., Chapiro, J., Paradis, V., Seraphin, T. P. & Kather, J. N. Artificial intelligence in liver diseases: improving diagnostics, prognostics and response prediction. *JHEPReport* **0**, (2022).
7. Brockmoeller, S. *et al.* Deep learning identifies inflamed fat as a risk factor for lymph node metastasis in early colorectal cancer. *J. Pathol.* **256**, 269–281 (2022).
8. Schrammen, P. L. *et al.* Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *J. Pathol.* (2021) doi:10.1002/path.5800.
9. Echle, A. *et al.* Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. *Gastroenterology* **159**, 1406–1416.e11 (2020).
10. Pallua, J. D., Brunner, A., Zelger, B., Schirmer, M. & Haybaeck, J. The future of pathology is digital. *Pathol. Res. Pract.* **216**, 153040 (2020).
11. Niazi, M. K. K., Parwani, A. V. & Gurcan, M. N. Digital pathology and artificial intelligence. *Lancet Oncol.* **20**, e253–e261 (2019).
12. Herrington, C. S., Poulsom, R. & Coates, P. J. Recent Advances in Pathology: the 2020 Annual Review Issue of The Journal of Pathology. *J. Pathol.* **250**, 475–479 (2020).
13. Kleppe, A. *et al.* Chromatin organisation and cancer prognosis: a pan-cancer study. *Lancet Oncol.* **19**, 356–369 (2018).
14. Courtiol, P. *et al.* Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
15. Eykholt, K. *et al.* Robust Physical-World Attacks on Deep Learning Models. *arXiv [cs.CR]* (2017).
16. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. & Mukhopadhyay, D. Adversarial Attacks and Defences: A Survey. *arXiv [cs.LG]* (2018).
17. Gordon, W. J. & Stern, A. D. Challenges and opportunities in software-driven medical devices. *Nat Biomed Eng* **3**, 493–497 (2019).
18. Finlayson, S. G. *et al.* Adversarial attacks on medical machine learning. *Science* **363**, 1287–1289 (2019).
19. Foote, A. *et al.* Now You See It, Now You Dont: Adversarial Vulnerabilities in Computational Pathology. *arXiv [eess.IV]* (2021).
20. Albawi, S., Mohammed, T. A. & Al-Zawi, S. Understanding of a convolutional neural network. in *2017 International Conference on Engineering and Technology (ICET)* 1–6 (2017).
21. Vaswani, A. *et al.* Attention Is All You Need. *arXiv [cs.CL]* (2017).
22. Tuli, S., Dasgupta, I., Grant, E. & Griffiths, T. L. Are Convolutional Neural Networks or Transformers more like human vision? *arXiv [cs.CV]* (2021).
23. Laleh, N. G. *et al.* Benchmarking artificial intelligence methods for end-to-end computational pathology. *bioRxiv* 2021.08.09.455633 (2021) doi:10.1101/2021.08.09.455633.
24. Chen, R. J. *et al.* Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 4015–4025 (2021).
25. Aldahdooh, A., Hamidouche, W. & Deforges, O. Reveal of Vision Transformers Robustness against Adversarial Attacks. *arXiv [cs.CV]* (2021).
26. Mahmood, K., Mahmood, R. & Van Dijk, M. On the robustness of vision transformers to adversarial

examples. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 7838–7847 (2021).
27. Shao, R., Shi, Z., Yi, J., Chen, P.-Y. & Hsieh, C.-J. On the adversarial robustness of visual transformers. *arXiv e-prints* arXiv–2103 (2021).
28. Norgeot, B. *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.* **26**, 1320–1324 (2020).
29. Dislich, B., Blaser, N., Berger, M. D., Gloor, B. & Langer, R. Preservation of Epstein-Barr virus status and mismatch repair protein status along the metastatic course of gastric cancer. *Histopathology* **76**, 740–747 (2020).
30. Macenko, M. *et al.* A method for normalizing histology slides for quantitative analysis. in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 1107–1110 (2009).
31. Kather, J. N. *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer* **1**, 789–799 (2020).
32. Muti, H. S. *et al.* Development and validation of deep learning classifiers to detect Epstein-Barr virus and microsatellite instability status in gastric cancer: a retrospective multicentre cohort study. *The Lancet Digital Health* vol. 3 e654–e664 (2021).
33. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).
34. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv [cs.CV]* (2020).
35. Han, T. *et al.* Advancing diagnostic performance and clinical usability of neural networks via adversarial training and dual batch normalization. *Nat. Commun.* **12**, 1–11 (2021).
36. Liu, Y., Mao, S., Mei, X., Yang, T. & Zhao, X. Sensitivity of Adversarial Perturbation in Fast Gradient Sign Method. in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)* 433–436 (2019).
37. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv [stat.ML]* (2014).
38. Kurakin, A., Goodfellow, I. & Bengio, S. Adversarial examples in the physical world. *arXiv [cs.CV]* (2016).
39. Madry, A., Makelov, A. & Schmidt, L. Towards Deep Learning Models Resistant to Adversarial Attacks.
40. Croce, F. & Hein, M. Minimally distorted Adversarial Examples with a Fast Adaptive Boundary Attack. in *Proceedings of the 37th International Conference on Machine Learning* (eds. Iii, H. D. & Singh, A.) vol. 119 2196–2205 (PMLR, 13--18 Jul 2020).
41. Andriushchenko, M., Croce, F., Flammarion, N. & Hein, M. Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search. in *Computer Vision – ECCV 2020* 484–501 (Springer International Publishing, 2020).
42. Wang, L., Zhang, Y. & Feng, J. On the Euclidean distance of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1334–1339 (2005).
43. Muti, H. S. *et al.* The Aachen Protocol for Deep Learning Histopathology: A hands-on guide for data preprocessing. (2020) doi:10.5281/ZENODO.3694994.
44. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering* 1–16 (2021).
45. Petrelli, F. *et al.* Prognostic value of diffuse versus intestinal histotype in patients with gastric cancer: a systematic review and meta-analysis. *Journal of Gastrointestinal Oncology* vol. 8 148–163 (2017).
46. Wang, K. *et al.* A cohort study and meta-analysis of the evidence for consideration of Lauren subtype when prescribing adjuvant or palliative chemotherapy for gastric cancer. *Ther. Adv. Med. Oncol.* **12**, 1758835920930359 (2020).
47. Ma, J., Shen, H., Kapesa, L. & Zeng, S. Lauren classification and individualized chemotherapy in gastric cancer. *Oncol. Lett.* **11**, 2959–2964 (2016).
48. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* **3**, 118 (2020).
49. Liu, S. & Cheng, B. Cyberattacks: Why, What, Who, and How. *IT Prof.* **11**, 14–21 (2009).
50. Bhojanapalli, S. *et al.* Understanding Robustness of Transformers for Image Classification. *arXiv*

[cs.CV] (2021).
51. Paul, S. & Chen, P.-Y. Vision Transformers are Robust Learners. *arXiv [cs.CV]* (2021).
52. Su, J., Vargas, D. V. & Sakurai, K. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Trans. Evol. Comput.* **23**, 828–841 (2019).
53. Fort, S. Pixels still beat text: Attacking the OpenAI CLIP model with text patches and adversarial pixel perturbations. *Stanislav Fort* https://stanislavfort.github.io/blog/OpenAI_CLIP_stickers_and_adversarial_examples/ (2021).
54. Dong, Y. *et al.* Boosting adversarial attacks with momentum. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 9185–9193 (2018).
55. Rao, C. *et al.* A Thorough Comparison Study on Adversarial Attacks and Defenses for Common Thorax Disease Classification in Chest X-rays. *arXiv [eess.IV]* (2020).
56. Brendel, W., Rauber, J. & Bethge, M. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. *arXiv [stat.ML]* (2017).
57. Bhagoji, A. N., He, W., Li, B. & Song, D. Practical black-box attacks on deep neural networks using efficient query mechanisms. in *Computer Vision – ECCV 2018* 158–174 (Springer International Publishing, 2018).