Adversarial Attacks Against Medical Deep Learning Systems

Samuel G. Finlayson
Department of Systems Biology
Harvard Medical School
Boston, MA 02130
samuel_finlayson@hms.harvard.edu

Isaac S. Kohane
Department of Biomedical Informatics
Harvard Medical School
Boston, MA 02130
isaac_kohane@hms.harvard.edu

Andrew L. Beam
Department of Biomedical Informatics
Harvard Medical School
Boston, MA 02130
andrew_beam@hms.harvard.edu

Abstract

The discovery of adversarial examples has raised concerns about the practical deployment of deep learning systems. In this paper, we argue that the field of *medicine* may be uniquely susceptible to adversarial attacks, both in terms of monetary incentives and technical vulnerability. To this end, we outline the healthcare economy and the incentives it creates for fraud, we extend adversarial attacks to three popular medical imaging tasks, and we provide concrete examples of how and why such attacks could be realistically carried out. For each of our representative medical deep learning classifiers, both white and black box attacks were both effective and human-imperceptible. We urge caution in employing deep learning systems in clinical settings, and encourage research into domain-specific defense strategies.

1 Introduction

Over the past six years, deep learning has transformed computer vision and has been implemented in scores of consumer-facing products. Many are excited that these approaches will continue to expand in scope and that new tools and products will be improved through the use of deep learning. One particularly exciting application area of deep learning has been in clinical applications. There are many recent high-profile examples of deep learning achieving parity with human physicians on tasks in radiology [18, 51], pathology [6], and opthalmology [22]. In some instances, the performance of these algorithms exceed the capabilities of most *individual* physicians in head-to-head comparisons. This has lead some to speculate that entire specialties in medical imaging, such as radiology and pathology, may be radically reshaped [24] or cease to exist entirely. Furthermore, on April 11, 2018, an important step was taken towards this future: the U.S. Food and Drug Administration announced the approval of the first computer vision algorithm that can be utilized for medical diagnosis without the input of a human clinician [1].

In parallel to this progress in medical deep learning, the discovery of so-called 'adversarial examples' has exposed vulnerabilities in even state-of-the-art learning systems [19]. Adversarial examples – inputs engineered to cause misclassification – have quickly become one of the most popular areas of research in the machine learning community [56, 39, 37, 47]. While much of the interest with adversarial examples has stemmed from their ability to shed light on possible limitations of current deep learning methods, adversarial examples have also received attention because of the cybersecurity threats they may pose for deploying these algorithms in both virtual and physical settings [47, 36, 28, 4, 9, 21]. Nevertheless, the literature has yet to thoroughly address the possibility of adversarial attacks in a medical context.

Given the enormous costs of healthcare in the US, it may seem prudent to take the expensive human 'out of the loop' and replace him or her with an extremely cheap and highly accurate deep learning algorithm. This seems especially tempting given a recent study's finding that physician and nursing pay is one of the key drivers of high costs in the US relative to other developed countries [44]. However, there is an under-appreciated downside to widespread automation of medical imaging tasks given the current vulnerabilities of these algorithms: If we seriously consider taking the human doctor completely 'out of the loop' (which now has legal sanction in at least one setting via the FDA, with many more to likely follow), we are forced to also consider how adversarial attacks may present new opportunities for fraud and harm. In fact, even with a human in the loop, any clinical system that leverages a machine learning algorithm for diagnosis, decision-making, or reimbursement could be manipulated with adversarial examples.

In this paper, we argue that healthcare is particularly vulnerable to adversarial attacks and that there are enormous incentives to motivate prospective bad actors to carry out these attacks. We extend previous results on adversarial examples to three medical deep learning systems modeled after the state of the art medical classifiers, and we provide perspective on the scope of potential for adversarial attacks in medicine. Because the healthcare system is complex and administrative processes can appear byzantine, it may be difficult to imagine how these attacks could be operationalized. As such, to ground the abstract potential for harm in actual use cases, we describe a future where many tasks have been fully automated by deep learning and give specific examples of the fraud that could be mediated by adversarial attacks. Our goal is to provide background on the distinct features of the medical pipeline that make adversarial attacks a threat, and also to demonstrate the practical feasibility of these attacks on real medical deep learning systems. We hope that increased awareness of the potential harm from adversarial examples in medicine encourage more in the machine learning community work towards solutions that will allow these techniques to be safely deployed within healthcare.

2 Adversarial examples

Adversarial examples are inputs to machine learning models that have been crafted to force the model to make a classification error. In a sense, this problem extends back in time at least as far as the spam filter, where systematic modifications to email such as 'good word attacks' or spelling modifications have long been employed to try to bypass the filters [31, 32, 13]. More recently adversarial examples were first discovered and described in the context of deep computer vision systems through the work of Szegedy et al. [56] and Goodfellow et al. [19]. Particularly intriguing in these early examples was the fact that adversarial examples could be crafted to be extremely effective despite being imperceptibly different from natural images to human eyes. In the years since, adversarial examples – visible and otherwise – have been shown to exist for a wide variety of classic and modern learning systems [46, 9, 15]. By the same token, adversarial examples have been extended to various other domains such as text and speech processing [25, 11]. For an interesting history of adversarial examples and methods used to combat them, see Biggio and Roli [7].

Figure 1 shows a high-level example of what one real and human-imperceptible adversarial example looks like in the context of melanoma detection. The image was constructed using the projected gradient descent (PGD) method developed by Kurakin et al. [28] and Madry et al. [34]. While imperceptibility is not necessarily a universal criterion for adversarial examples, the existence of imperceptibly small adversarial perturbations demonstrates that the model is in fact being fooled by the attacker and that carefully designed attacks could be undetectable even with a human in the loop to perform visual inspection.

Adversarial examples are generally thought to arise from the piecewise linear components of even complex nonlinear models Goodfellow et al. [19]. They are not random, they are not due to overfitting or incomplete model training, they occupy only a comparatively small subspace of the feature landscape, they are robust to random noise, and they have been shown to transfer in many cases from one model to another [46, 57]. Furthermore, in addition to executing many successful attacks in

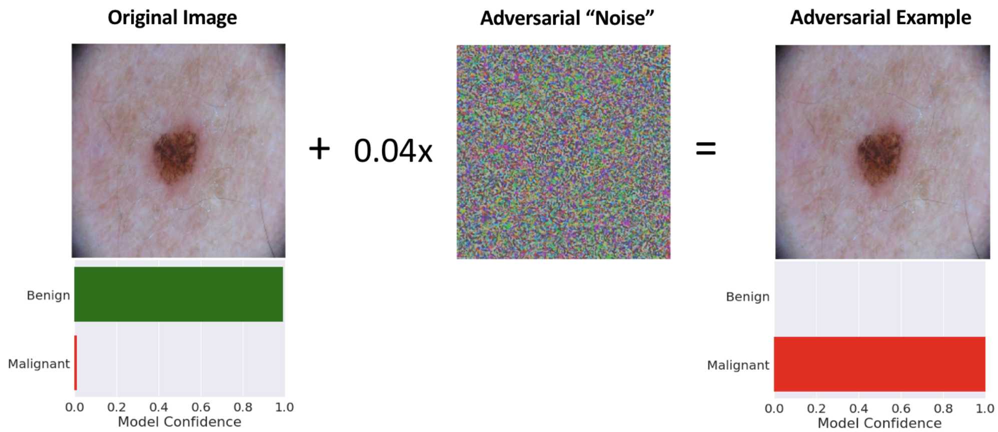

Figure 1: Overview of an adversarial example: The addition of a carefully crafted perturbation converts an image that the model correctly classifies as benign into an image that the network is 100% confident is malignant.

purely virtual settings, researchers in the past several years have also demonstrated that adversarial attacks can generalize to physical world settings [28, 17, 4, 9].

One natural question raised by the existence of adversarial examples is to what extent and in what forms they constitute a viable risk for harm in real-world AI settings. Many authors have discussed the feasibility of and possible motivations for adversarial attacks on certain real-world systems such as self-driving cars [17, 12, 33]. However, to our knowledge, the AI literature is yet to thoroughly discuss the possibility of adversarial attacks on medical systems.

3 Features of the healthcare system that favor adversarial attacks

3.1 Background on the healthcare economy and possible incentives for fraud

The healthcare economy is huge and fraud is already pervasive. Though most healthcare workers have good intentions, the incentives for defrauding the system are enormous. The United States spent approximately $3.3 trillion (17.8% of GDP) on healthcare in 2016 [44], and healthcare is projected to represent 1/5 of the US economy by 2025. Given the vast sums of money represented by the healthcare economy, inevitably some seek to profit by committing fraud against the healthcare system. Medical fraud is estimated to cost hundreds of billions of dollars each year, and one study estimated this number to be as high as $272 billion in 2011 [23]. Fraud can be committed at all levels of the healthcare system by actors with varying degrees of leverage and power. Large institutions engage in fraud by systemically inflating costs for services to increase revenue [48, 53]. Likewise, it has been found that some individual physicians routinely bill for the highest allowable amount over 90% of the time [43]. Thus, in the existing system fraud committed both by large institutions and also by individual actors.

Every data point represents real dollars. Due to amount of money involved with the delivery of healthcare, complex book-keeping systems have been created to facilitate billing and reimbursement, and are routinely used by the healthcare 'payers' (private or public insurers). In fact, most of the data generated by the healthcare system in the electronic healthcare record (EHR) is created in large part to justify payments from payers to 'providers' (hospitals and physicians). To facilitate this process, diagnostic, therapeutic, and procedural coding schema have been developed to describe – with very fine granularity[1] – a given patient's diagnoses and associated treatments. One of the most popular coding schemas is the *International Classification of Diseases, 10th revision* [42] (ICD-10) which contains codes representing over 70,000 different diagnoses. Since individual codes are often associated with a specific monetary value of reimbursement, even small differences in

[1]See for instance the ICD-10 code V97.33XD which represents the following diagnosis: 'Sucked into jet engine, subsequent encounter'

coding practices can amount to millions of dollars of reimbursement for a hospital or medical practice. As a result, both providers and payers invest billions of dollars per year on IT infrastructure and support staff to try to optimize their respective components of the billing process.

In an effort to increase revenue, some providers have engaged in the practice of 'upcoding' diagnoses or procedures by selecting the codes which will allow them to bill for the highest amount. For their part, insurance companies seek to minimize total expenditure by identifying the minimal medically indicated healthcare spending. In extreme cases, refusal by the insurance company to pay for a specific therapy renders it effectively impossible for the patient to receive the drug or procedure, regardless of the opinion of the physician. Unsurprisingly, the establishment of the correct code (which will dictate the billable amount) for a given patient can become both heated and expensive in practice, and can even lead to litigation. To ensure consistency and justifiability, insurance companies will often demand specific gold standard metrics as proof of diagnosis before reimbursing a given medical claim.

Algorithms will likely make reimbursement decisions in the future. Given these dynamics, it is seemingly inevitable that insurance companies will begin to require that 'unbiased' approaches (such as deep learning) confirm certain diagnoses and justify resultant reimbursement. If and when this occurs, the ability to undetectably influence (either as a provider or as a payer) the outputs of trusted and otherwise unbiased diagnostic systems would result in the ability to influence the movement of billions of dollars through the healthcare economy.

Incentives for medical adversarial fraud extend beyond clinical practice settings. While we've focused the above discussion on the lesser-known incentives for fraud in clinical practice settings, it should also be noted that diagnostic algorithms are likely to form a key aspect of component of the broader biomedical industry. For example, the pharmaceutical industry increasingly relies upon imaging analyses to demonstrate the efficacy of drugs in research and clinical trials [49], with each government decision to approve or reject a drug often amounting to billions of dollars in profits gained or lost. Given that at least one study has found that for-profit companies are much more likely to report a positive result than non-profit entities when conducting a clinical trial, some have expressed concerns that certain pharmaceutical companies may already engage in practices to try to influence drug trial results [52].

Incentives exist for adversarial attacks against non-imaging classifiers. This paper focuses on adversarial examples against medical *image* classifiers, in part because this has been the predominant focus of most adversarial example research. However, it should be noted that the above incentives apply equally to other types of medical machine learning algorithms, such as diagnostics that employ natural language processing or signal processing algorithms. These types of classifiers have been developed for clinical diagnostics and also shown to be vulnerable to adversarial attacks [25, 30, 2, 11]. The task of detecting medical fraud, usually based on analyzing insurance claims datasets and electronic medical records, has also been extensively studied and implemented as a use-case for machine learning in healthcare [38]. Designing precise alterations to medical records or billing patterns to evade detection by such systems is another clear use-case for adversarial examples.

3.2 Distinctive sources of vulnerability to medical adversarial attacks among medical deep learning systems

The field of medical imaging has technical features that could make it uniquely vulnerable to adversarial attacks. We provide a sketch of these vulnerabilities below.

Ground truth is often ambiguous. Compared to most common image classification tasks, the ground truth in medical imaging is often controversial, with even specialty radiologists disagreeing on well defined tasks [40, 29, 8] As such, if attackers selectively perturb images for which it is difficult to establish the true diagnosis, they can make it extremely difficult to detect their influence through even expert human review.

Medical imaging is highly standardized. With very few exceptions, the results of medical imaging tests are not only static, but also extremely standardized in terms of both positioning and exposure [55]. Medical adversarial attacks thus do not need to meet the same standards of invariance to lighting and positional changes as attacks on other real-world systems such as self-driving cars. This is potentially important, as some have argued that dynamic viewing conditions imply that 'there is a good prospect that adversarial examples can be reduced to a curiosity with little practical impact' [33].

Commodity network architectures are often used. Nearly all of the most successful published methods in medical computer vision have consisted of the same fundamental architecture: one of a small set of pretained imagenet model that was fine-tuned to the specific task [22, 58, 16]. This lack of architectural diversity could make it easier for potential attackers to build transferable attacks against medical systems. By the same token, given the importance of peer review and publication in validating and approving medical diagnostics, it is likely that at least the architectures of most medical AI models will be public for the sake of transparency, allowing for more informed adversarial attacks.

Medical data interchange is limited and balkanized. Five electronic health record (EHR) vendors constitute about half of the market and hundreds of others serve the other half. Even within an EHR vendor, data sharing is spotty and the terminologies and their semantics vary considerably from one implementation to another. On the one hand this means that there are no universally shared mechanisms for authentication, verification of message integrity, data quality metrics, nor mechanisms for automated oversight. On the other hand this allows healthcare providers to customize their EHR's, billing and other information technology systems in ways that are opaque to most external auditors using one-size-fits-all tools and methods.

Hospital infrastructure is very hard to update. Medical software is often implemented within monolithic enterprise-wide proprietery software systems making updates, revisions and fixes expensive and time-consuming. For context, consider the coding dictionaries used to classify patients' diseases, the International Classification of Disease (ICD) system. As recently as 2013, most hospitals were operating using the ninth edition of this coding scheme, published in 1978, despite the fact that a revised version (ICD-10) was published in 1990. All told, the conversion to the ICD-10 coding scheme has been estimated to cost major health centers up to $20 million *each* and require up to 15 years [54]. Others decided in the early 2000s that it would be preferable to skip the 1990 schema entirely and wait for ICD-11, despite the fact that its release wasn't scheduled until 2018. Thus, vulnerabilities present in medical software are likely to persist for years due to the difficulty and expense involved with update hospital infrastructure.

For further context in this regard, one can consider that the primary means of transferring text records between hospitals continues to be the *fax machine*. When it comes to images, a patient trying to complete the simple task of obtaining images to send to a new physician will in many cases have to write a letter and wait weeks for a manual image export to a disk drive. This is not an industry that could roll out complex image pipeline updates overnight.

Medicine contains a mix of technical and non-technical workers. Compared to many other industries, medicine is extremely interdisciplinary and mostly comprised of members who lack a strong computational or statistical training background. For example, in the case of self-driving cars, the teams developing computer vision systems are likely to be led and staffed primarily by engineers, even if not computer scientists. In contrast, since the clinical usability of medical imaging systems is also extremely important, hospitals are likely to lean heavily on physician-researchers in developing these systems, who tend to lack robust computational expertise [35].

There are many potential adversaries. The medical imaging pipeline has many potential attackers and is thus vulnerable at many different stages. While in theory one could devise elaborate image verification schemes throughout the image processing pipeline to try to guard against attacks, this would be extremely costly and difficult to implement in practice.

4 Demonstration of adversarial attacks on medical deep learning systems

4.1 Construction of medical classification models

To demonstrate the feasibility of adversarial attacks on medical deep learning systems, we first developed baseline models to classify referable diabetic retinopathy from retinal fundoscopy (similar to Gulshan et al. [22]), pneumothorax from chest-xray (similar to Wang et al. [58] and Rajpurkar et al. [51]), and melanoma from dermoscopic photographs (similar to Esteva et al. [16]). The decision to build models for these particular tasks was made both due to public data availability, as well as the fact that they represent three of the most highly visible successes for medical deep learning.

Inputs	Accuracy	AUC	Sensitivity	Specificity
Fundoscopy	91%	0.91	85%	80%
Chest X-Ray	95%	0.94	90%	82%
Dermoscopy	88%	0.86	80%	73%

Table 1: Results of baseline medical AI models on the validation set.

All of our models were trained on publically available data. For diabetic retinopathy, this was the Kaggle Diabetic Retinopathy dataset. The key distinction with the Kaggle dataset, however, was that we were seeking to predict *referable* (grade 2 or worse) diabetic retinopathy in accordance with Gulshan et al. [22] rather than predicting the retinopathy grade itself as was the case in the Kaggle competition. As such, the training and test sets from the kaggle competition were merged, relabeled using their provided grades, and split by patient into training and test sets with probability 0.88/0.12. For the chest x-rays, we used the ChestX-Ray14 dataset described by Wang et al. [58]. We identified cases and controls by selecting images whose labeled contained 'pneumothorax' or 'no finding', respectively, and additionally excluded from our control group any images from patients who had received both labels. We then split by patient into training and test sets with probability 0.85/0.15. For melanoma, we downloaded images labeled as benign or malignant melanocytic lesions from the International Skin Imaging Collaboration website (isic-archive.com), splitting again into training and test sets with probability 0.85/0.15.

As in the case of all three of the original papers that inspired these models, we built our classifiers by fine-tuning a pretrained ImageNet model. For convenience and consistency, we chose to build each of our networks using a pretrained ResNet-50 model, fine-tuned in Keras using stochastic gradient descent with a learning rate of 1E-3 and Momentum of 0.9. Data was augmented using 45° rotation and horizontal flipping chest x-ray images, and with 360° rotation, vertical and horizontal flipping, and Mixup for fundoscopy and dermoscopy images. In all three cases, these settings provided respectable performance and we therefore didn't perform dedicated hyperparameter optimization.

The resulting baseline classifiers had strong performance, presented in Table 4.1.

4.2 Adversarial attacks

To demonstrate the vulnerability of our models to adversarial attacks, we followed the white and black box PGD attack strategies recently published in Madry et al. [34], which were also selected as the gold-standard baselines in Kannan et al. [26].

The PGD attack (see Madry et al. [34] and also Kurakin et al. [28]) is an iterative extension of the canonical fast gradient sign method (FGSM) attack developed by Goodfellow et al. [19]. In the PGD attack, given input $x \in \mathbb{R}^d$, loss function $L(\theta, x, y)$, and a set of allowed perturbations $\mathcal{S} \subseteq \mathbb{R}^d$ (most commonly the ℓ_∞ ball around x), one can perform a projected gradient descent on the negative loss function:

$$x^{t+1} = \Pi_{x+\mathcal{S}}(x^t + \epsilon \text{sgn}(\nabla_x L(\theta, x, y)))$$

in order to identify an optimal perturbation. We implemented the PGD attack using the library Cleverhans, conducting 20 iterations with hyperparameter $\epsilon = 0.02$ (ϵ corresponds to the maximum permitted ℓ_∞ norm of the perturbation) [45].

White box attacks were implemented by running PGD directly on the victim model itself. As in Madry et al. [34] and Kannan et al. [26], black box attacks were performed by running the PGD attack against an independently trained model with the same architecture and transferring the resultant adversarial examples to the victim. While this form of black box attack is slightly 'gray' due to the fact that we knowingly duplicate the overall architecture of the model, the attack makes no use of any weights or predictions generated by the victim model itself. Further, given the standard of publishing and publicly vetting medical systems prior to approval, it is entirely sensible to assume that many or most attackers will have knowledge of the architecture of the systems they attack.

Inputs	Accuracy	AUC	Avg. Conf.
Fundoscopy			
Clean	91.0%	0.910	90.4%
White Box	0.00%	0.000	100.0%
Black Box	0.01%	0.002	90.9%
Chest X-Ray			
Clean	94.9%	0.937	96.1%
White Box	0.00%	0.000	100.0%
Black Box	15.1%	0.014	92.6%
Dermoscopy			
Clean	87.6%	0.858	94.1%
White Box	0.00%	0.000	100.0%
Black Box	37.9%	0.071	92.0%

Table 2: Results of medical deep learning models on clean test set data, white box, and black box attacks.

The results of our attacks are depicted in Table 4.2 and in Figure 2. Unsurprisingly, they were effective against all three systems. Additional examples (along with the adversarial noise) can be found in the appendix. Code can be found at [https://github.com/sgfin/adversarial-medicine]

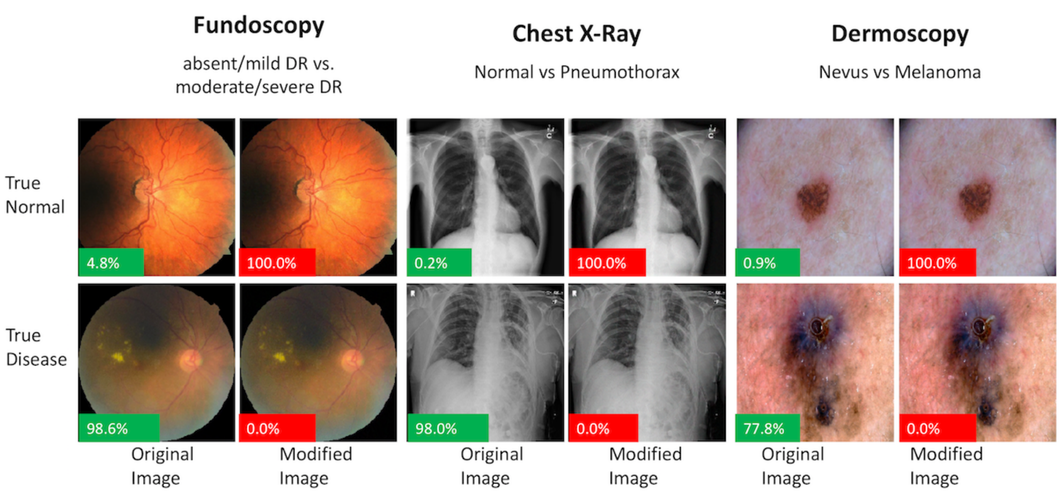

Figure 2: Characteristic results of adversarial example generation. The percentage displayed on the bottom left of each image represents the probability that the model assigns that image of being diseased. Green = Model is correct on that image. Red = Model is incorrect. As can be seen, in each case, human imperceptible changes were sufficient to make the classifier 100% confident in the wrong classification.

5 Discussion

We now discuss how someone might perform adversarial attacks against the systems developed in previous section under a realistic set of conditions. For the purposes of illustration, consider a scenario where these systems have been subjected to extensive testing and validation and are now clinically deployed. These systems would function much like laboratory tests do now and provide confirmation of suspected diagnoses. In some instances, an insurance company may require a confirmatory diagnosis from one of these systems in order for a reimbursement to be made. We provide the example below to show that in many instances there is both the *opportunity* and *incentive* for someone to use an adversarial example to defraud the healthcare system.

5.1 Hypothetical examples

Adversarial examples in dermatology: Dermatology in the US operates under a 'fee for service' model wherein a physician or practice is paid for the services or procedures they perform for the patient. Under this model, dermatologists are incentivized to perform as many procedures as possible, as their revenue is directly tied to the amount of procedures they perform. This has caused some dermatologists to perform a huge number of uneccessary procedures to increase revenue. Consider the following real example of fraud committed by a dermatologist in Florida, quoted from [53]:

> In Florida, a dermatologist was sentenced to 22 years in prison, paid $3.7 million in restitution, forfeited an addition $3.7 million, and paid a $25,000 fine for performing 3,086 medically unnecessary surgeries on 865 Medicare beneficiaries.

To combat fraud and unnecessary procedures such as this, an insurance company could require that a deep learning system (e.g. the one from Section 4) analyzes all dermoscopy images to confirm that surgery is necessary. In this scenario, a bad actor could add adversarial noise to images to ensure that the deep learning model always gives the diagnosis that he or she desires. Furthermore, they could add this noise to 'boderline' cases which would render the attack nearly impossible to detect by human review. Thus, a bad actor like the Florida dermatologist from [53] could use the deep learning algorithm to provide cover for his fraudulent operation and continue to defraud the system in perpetuity.

Adversarial examples in radiology. Chest X-rays provide a common screening test for dozens of diseases, and a positive chest X-ray result is often used to justify more heavily reimbursed procedures such as biopsies, CT or MR imaging, or surgical resection. As such, one could imagine many scenarios arising around chest X-rays that are directly analogous to the melanoma detection situation described above. However, we'd like to highlight a second class of incentives for fraud, namely those from the pharmaceutical industry: As discussed above, billions of dollars can hinge on a positive result for a drug undergoing a clinical trial. Thus, there are huge incentives to produce a positive result when testing a new drug. To reduce the ability of entities that run clinical trials to 'put their foot on the scale', the FDA could require that trial endpoints, such as tumor burden in chest imaging, be evaluated by a deep learning system such as the one from Section 4. By applying undetectable adversarial perturbations to the images, a company running a trial could effectively guarantee a positive trial result with respect to this endpoint.

Adversarial examples in ophthalmology. Providers and pharmaceutical companies are not the only organizations that could be incentivized to employ adversarial manipulation. Often the entities who pay for healthcare (such as private or public insurers) wish to curtail the utilization rates of certain procedures to reduce costs. However, there are often guidelines from government agencies (such as the Centers for Medicare and Medicaid Services) that specify diagnostic criteria which if present dictate that certain procedures must be covered. One such criterion could be that any patient with a confirmed diabetic retinopathy diagnosis from a deep learning system such as the one from Section 4 must have the resulting vitrectomy surgery covered by their insurer. Even though the insurer has no ability to control the policy, they could still control the rate of surgeries by using adversarial examples. By applying adversarial noise to mildly positive images, they could control the rate of positive diagnoses, reducing the number procedures without explicitly making changes to the underlying policy.

5.2 Possible defenses

Given the range of possible perverse incentives, we question if and precisely how fully automated medical imaging systems can be trusted in the absence of robust assurances against adversarial attacks. In this light, we briefly highlight two classes of defenses that could be developed to prevent adversarial attacks.

Algorithmic defenses against adversarial examples remain an extremely open and challenging problem, with recent state-of-the-art defenses on ImageNet still achieving only 27.9% and 46.7% top-1 accuracy for white- and black-box PGD attacks, respectively, as of March 2018 [26]. We defer a full discussion of the extremely rapidly evolving field of adversarial defenses to the primary literature [26, 10, 19, 27, 34, 46, 56, 57, 59, 47, 27, 50, 14]. Unfortunately, despite the explosive emergence of defense strategies, there does not appear to be an easy algorithmic fix for the adversarial problem available in the short term. For example, one recent analysis investigated a series of promising methods that relied on gradient obfuscation, and demonstrated that they could be quickly broken [3]. Despite this, we also note that principled approaches to adversarial robustness are beginning to show promise. For example, several papers have demonstrated what appears to be both high accuracy and strong adversarial robustness on smaller datasets such as MNIST, [34, 26], and there have also been several results including theoretical *guarantees* of adversarial robustness, albeit on small datasets and/or with still-insufficient accuracy [27, 50, 14]. Generalized attempts at algorithmic robustness are promising, but have yet to provide methods that can demonstrate high levels of both accuracy and adversarial robustness at ImageNet scale. Of note, application-specific defenses such as dataset-specific image preprocessing have also been shown to be highly effective on some datasets [20]. In this light, the standardized image capture procedures in biomedical imaging could provide an opportunity for defenders against adversarial attacks, if domain-specific defense strategies are discovered and applied. We feel that this is an important and promising area of future research.

Infrastructural defenses against clinical adversarial attacks include methods deployed to prevent potential bad actors from altering medical images – or at least make it easier to confirm image tampering if adversarial examples are suspected. For example, imaging devices could immediately store a hashed version of any image they generate, which could subsequently be used as a reference. Likewise, raw clinical images could be processed and analyzed on a third-party system to prevent any possible systemic manipulation by payers or providers. This family of approaches to standardized best practices is reminiscent of the system of Clinical Laboratory Improvement Amendments (CLIA), a set of federal policies that regulates the process by which clinical laboratory samples are handled and analyzed in the United States [41]. Given that algorithmic defenses against adversarial attacks are still very much an area of research, we feel that infrastructural defenses should be strongly considered for all medical AI systems that could carry incentives for adversarial attacks.

6 Conclusion

The prospect of improving healthcare and medicine with the use of deep learning is truly exciting. There is reasonable cause for optimism that these technologies can improve outcomes and reduce costs, if judiciously implemented [5]. In this light, it is unsurprising that dozens of private companies and large health centers have initiated efforts to deploy deep learning classifiers in clinical practice settings. As such efforts continue to develop, it seems inevitable that medical deep learning algorithms will become entrenched in the already multi-billion dollar medical information technology industry. However, the massive scale of the healthcare economy brings with it significant opportunity and incentive for fraudulent behavior.

In this work, we have outlined the systemic and technological reasons that cause adversarial examples to pose a disproportionately large threat in the medical domain. We have also demonstrated what we believe to the first examples of an adversarial attack being executed on medical systems. We hope that our results help facilitate a discussion on the threat of adversarial examples among both computer scientists and medical professionals. For machine learning researchers, we recommend research into infrastructural and algorithmic solutions designed to guarantee that attacks are infeasible or at least can be retroactively identified. For medical providers, payers, and policy makers, we hope that these practical examples can motivate a meaningful discussion into how precisely these algorithms should be incorporated into the clinical ecosystem despite their current vulnerability to such attacks.

Acknowledgements

The authors would like to thank Aleksander Madry and Dimitris Tsipras for a helpful review of our manuscript. In addition, SGF was supported by training grants T32GM007753 and T15LM007092; the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

References

[1] 2018. Press Announcements > FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. `https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357.htm`. (11 April 2018).

[2] Mohamad M Al Rahhal, Yakoub Bazi, Haikel AlHichri, Naif Alajlan, Farid Melgani, and Ronald R Yager. 2016. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences* 345 (2016), 340–354.

[3] Anish Athalye, Nicholas Carlini, and David Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420* (2018).

[4] Anish Athalye and Ilya Sutskever. 2017. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397* (2017).

[5] Andrew L Beam and Isaac S Kohane. 2016. Translating artificial intelligence into clinical care. *Jama* 316, 22 (2016), 2368–2369.

[6] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM van der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, and others. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* 318, 22 (2017), 2199–2210.

[7] Battista Biggio and Fabio Roli. 2017. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. *arXiv preprint arXiv:1712.03141* (2017).

[8] Charlotte L Brouwer, Roel JHM Steenbakkers, Edwin van den Heuvel, Joop C Duppen, Arash Navran, Henk P Bijl, Olga Chouvalova, Fred R Burlage, Harm Meertens, Johannes A Langendijk, and others. 2012. 3D variation in delineation of head and neck organs at risk. *Radiation Oncology* 7, 1 (2012), 32.

[9] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665* (2017).

[10] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. 2018. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*.

[11] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden Voice Commands.. In *USENIX Security Symposium*. 513–530.

[12] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 39–57.

[13] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, and others. 2004. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 99–108.

[14] Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy Mann, and Pushmeet Kohli. 2018. A Dual Approach to Scalable Verification of Deep Networks. *arXiv preprint arXiv:1803.06567* (2018).

[15] Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. 2018. Adversarial Examples that Fool both Human and Computer Vision. *arXiv preprint arXiv:1802.08195* (2018).

[16] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115.

[17] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. 2017. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945* (2017).

[18] William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P Bradley, and Lyle J Palmer. 2017. Detecting hip fractures with radiologist-level performance using deep neural networks. *arXiv preprint arXiv:1711.06504* (2017).

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

[20] Abigail Graese, Andras Rozsa, and Terrance E Boult. 2016. Assessing threat of adversarial examples on deep neural networks. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE, 69–74.

[21] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. 2017. Adversarial examples for malware detection. In *European Symposium on Research in Computer Security*. Springer, 62–79.

[22] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, and others. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316, 22 (2016), 2402–2410.

[23] Anita Jain, Samiran Nundy, and Kamran Abbasi. 2014. Corruption: medicine's dirty open secret. (2014).

[24] Saurabh Jha and Eric J Topol. 2016. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *Jama* 316, 22 (2016), 2353–2354.

[25] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328* (2017).

[26] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. 2018. Adversarial Logit Pairing. *arXiv preprint arXiv:1803.06373* (2018).

[27] J Zico Kolter and Eric Wong. 2017. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851* (2017).

[28] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* (2016).

[29] X Allen Li, An Tai, Douglas W Arthur, Thomas A Buchholz, Shannon Macdonald, Lawrence B Marks, Jean M Moran, Lori J Pierce, Rachel Rabinovitch, Alphonse Taghian, and others. 2009. Variability of target and normal structure delineation for breast cancer radiotherapy: an RTOG Multi-Institutional and Multiobserver Study. *International Journal of Radiation Oncology Biology Physics* 73, 3 (2009), 944–951.

[30] Znaonui Liang, Gang Zhang, Jimmy Xiangji Huang, and Qmming Vivian Hu. 2014. Deep learning for healthcare decision making with EMRs. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, 556–559.

[31] Daniel Lowd and Christopher Meek. 2005. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 641–647.

[32] Daniel Lowd and Christopher Meek. 2005. Good Word Attacks on Statistical Spam Filters.. In *CEAS*, Vol. 2005.

[33] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. 2017. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint arXiv:1707.03501* (2017).

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[35] Arjun K Manrai, Gaurav Bhatia, Judith Strymish, Isaac S Kohane, and Sachin H Jain. 2014. Medicine's uncomfortable relationship with math: calculating positive predictive value. *JAMA internal medicine* 174, 6 (2014), 991–993.

[36] Marco Melis, Ambra Demontis, Battista Biggio, Gavin Brown, Giorgio Fumera, and Fabio Roli. 2017. Is deep learning safe for robot vision? adversarial examples against the icub humanoid. *arXiv preprint arXiv:1708.06939* (2017).

[37] Seyed Mohsen Moosavi Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[38] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. 2015. Deep learning applications and challenges in big data analytics. *Journal of Big Data* 2, 1 (2015), 1.

[39] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 427–436.

[40] CF Njeh. 2008. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *Journal of medical physics/Association of Medical Physicists of India* 33, 4 (2008), 136.

[41] US Department of Health, Human Services, and others. 1992. Medicare, Medicaid and CLIA programs: regulations implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA). Final rule. *Fed Regist* 57, 40 (1992), 7002–7186.

[42] World Health Organization and others. 1990. International Classification of Diseases (10th revision). *Geneva. Switzerland. WHO* (1990).

[43] Charles Ornstein and Ryann Grochowski-Jones. 2018. Top Billing: Meet the Docs who Charge Medicare Top Dollar for Office Visits — ProPublica. (2018). `https://www.propublica.org/article/billing-to-the-max-docs-charge-medicare-top-rate-for-office-visits`

[44] Irene Papanicolas, Liana R Woskie, and Ashish K Jha. 2018. Health care spending in the United States and other high-income countries. *JAMA* 319, 10 (2018), 1024–1039.

[45] Nicolas Papernot, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Fartash Faghri, Alexander Matyasko, Karen Hambardzumyan, Yi-Lin Juang, Alexey Kurakin, Ryan Sheatsley, and others. 2016. cleverhans v2. 0.0: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768* (2016).

[46] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277* (2016).

[47] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 582–597.

[48] Kalb PE. 1999. Health care fraud and abuse. *JAMA* 282, 12 (1999), 1163–1168. DOI: `http://dx.doi.org/10.1001/jama.282.12.1163`

[49] Homer H Pien, Alan J Fischman, James H Thrall, and A Gregory Sorensen. 2005. Using imaging biomarkers to accelerate drug development and clinical trials. *Drug discovery today* 10, 4 (2005), 259–266.

[50] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. 2018. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344* (2018).

[51] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, and others. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225* (2017).

[52] Paul M Ridker and Jose Torres. 2006. Reported outcomes in major cardiovascular clinical trials funded by for-profit and not-for-profit organizations: 2000-2005. *Jama* 295, 19 (2006), 2270–2274.

[53] William J Rudman, John S Eberhardt, William Pierce, and Susan Hart-Hester. 2009. Healthcare fraud and abuse. *Perspectives in Health Information Management/AHIMA, American Health Information Management Association* 6, Fall (2009).

[54] Tekla B Sanders, Felicia M Bowens, William Pierce, Bridgette Stasher-Booker, Erica Q Thompson, and Warren A Jones. 2012. The road to ICD-10-CM/PCS implementation: forecasting the transition for providers, payers, and other healthcare organizations. *Perspectives in health information management/AHIMA, American Health Information Management Association* 9, Winter (2012).

[55] George Simon. 1971. *Principles of chest X-ray diagnosis*. Butterworths.

[56] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. *arXiv* (2014), 1–10. DOI:http://dx.doi.org/10.1021/ct2009208

[57] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2017. The Space of Transferable Adversarial Examples. *arXiv* (2017), 1–15. http://arxiv.org/abs/1704.03453

[58] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3462–3471.

[59] Weilin Xu, David Evans, and Yanjun Qi. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155* (2017).

Appendix

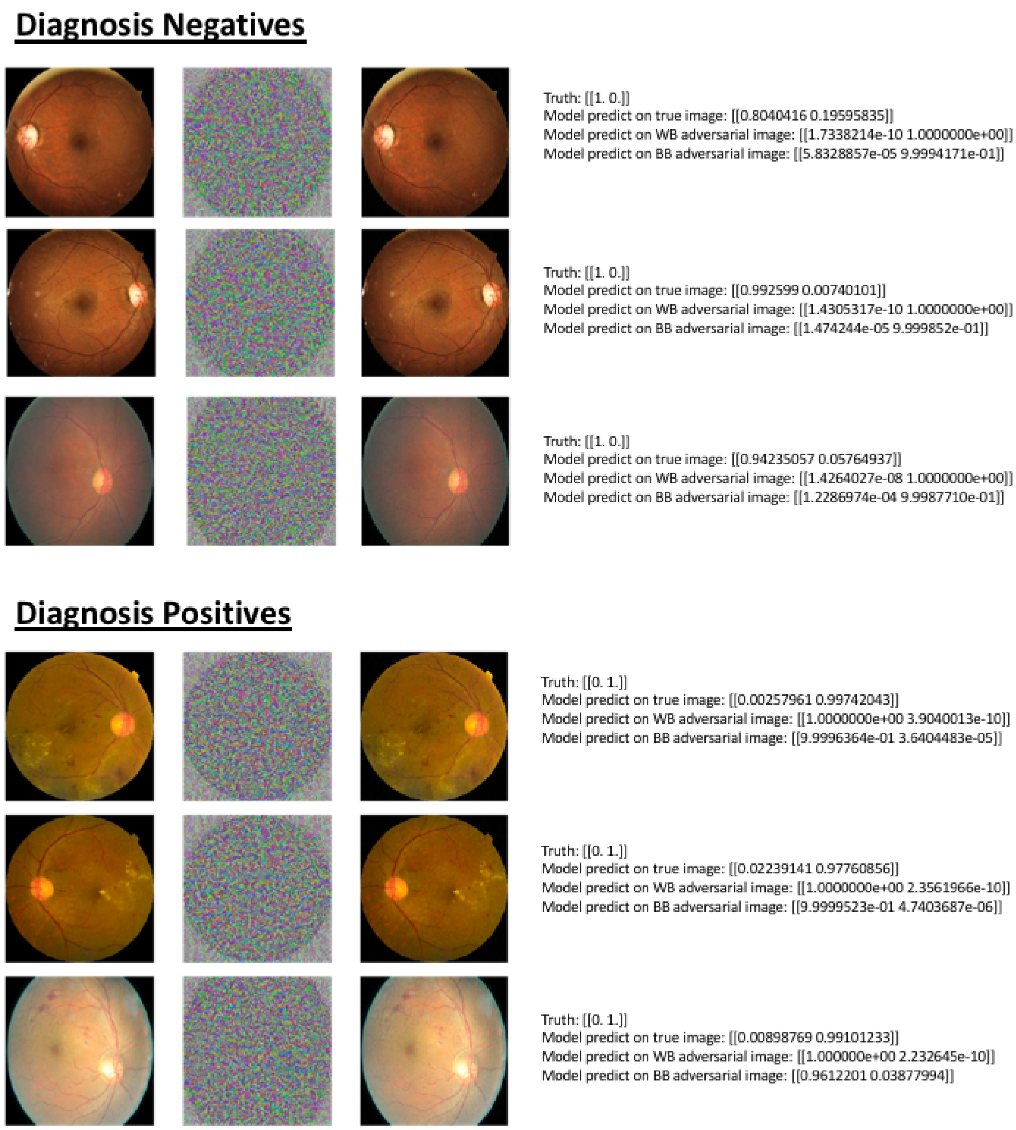

Figure 3: Additional examples of retinal fundus images, adversarial perturbations, and resultant adversarial examples. Noise displayed is the white box noise.

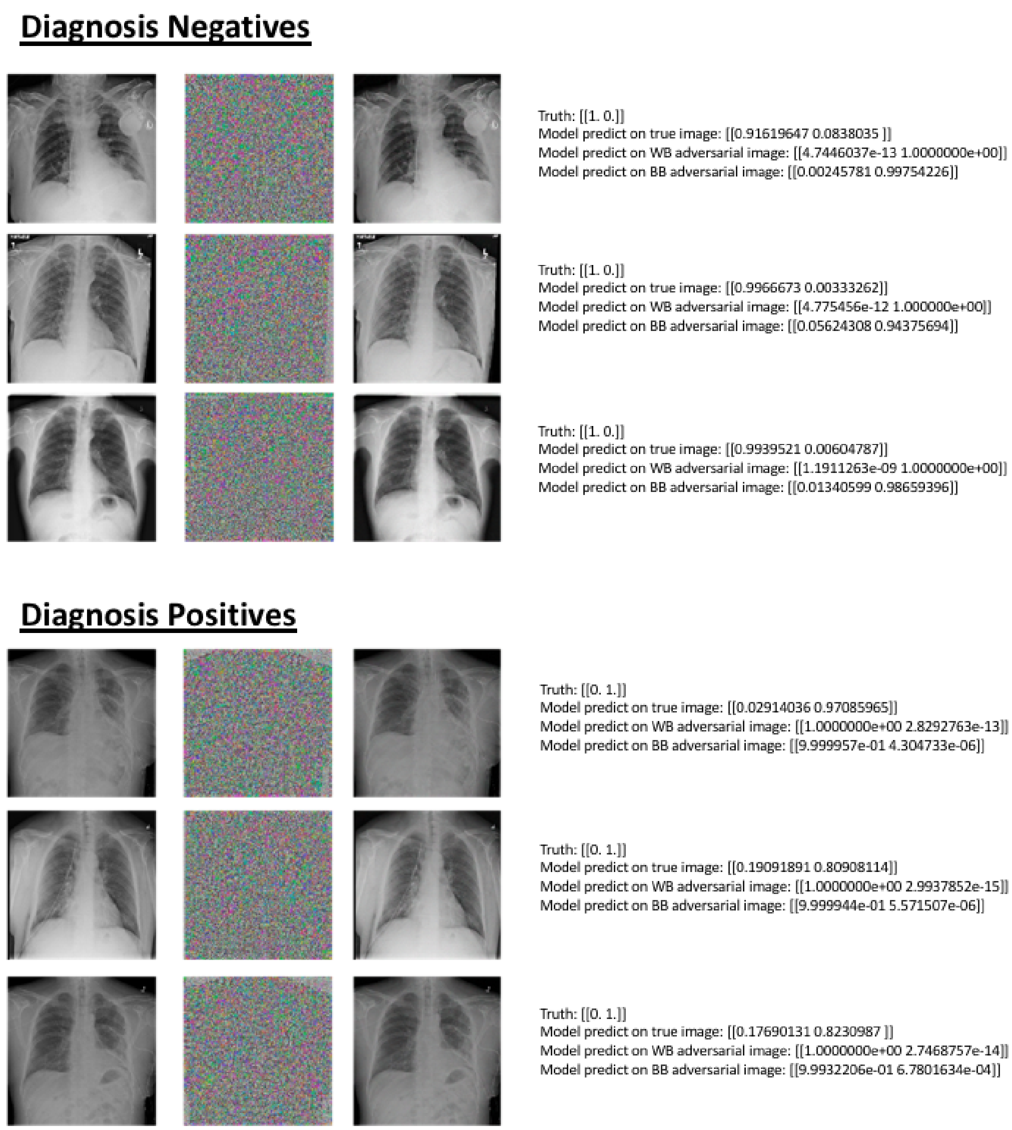

Figure 4: Additional examples of chest x-ray images, adversarial perturbations, and resultant adversarial examples. Noise displayed is the white box noise.

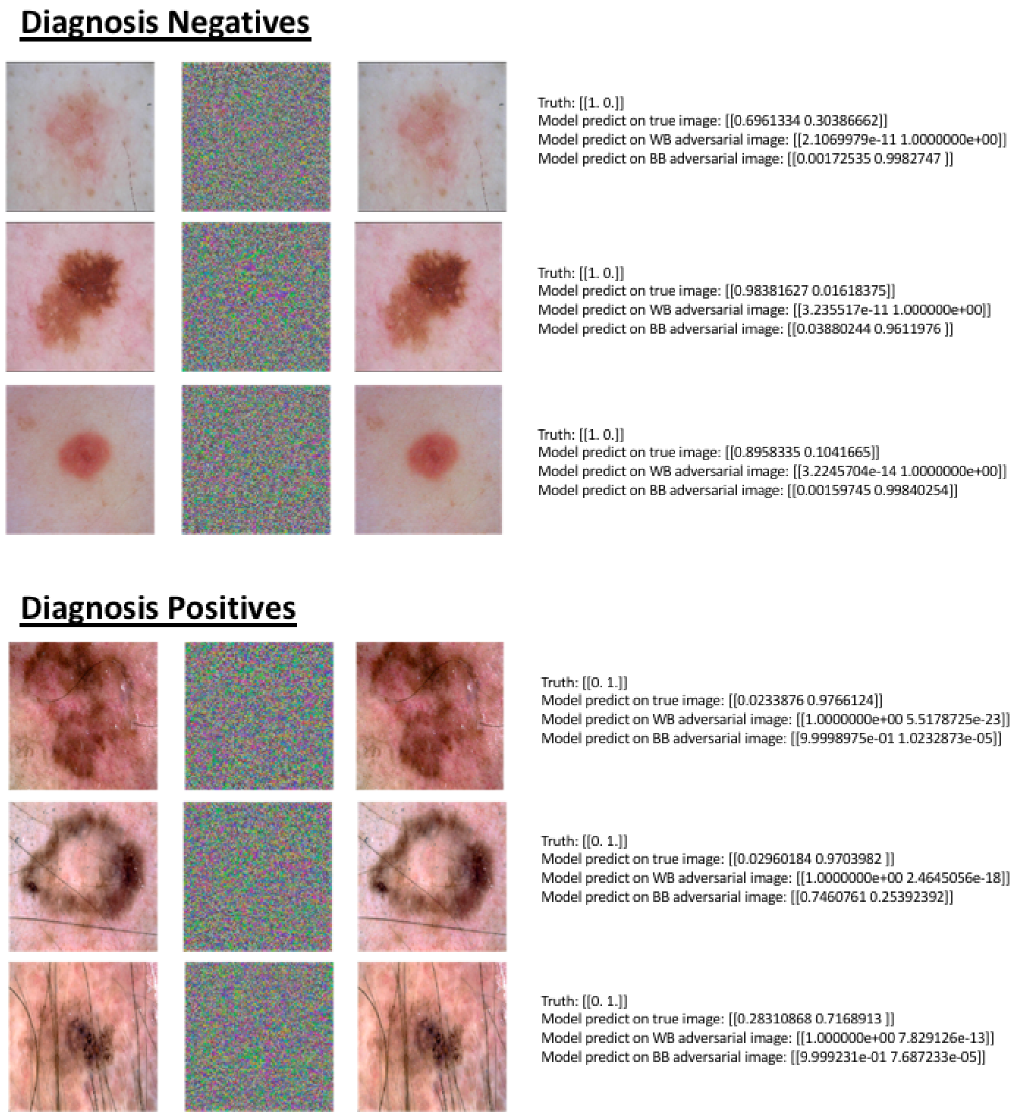

Figure 5: Additional examples of dermoscopy images, adversarial perturbations, and resultant adversarial examples. Noise displayed is the white box noise.