Physical Attacks in Dermoscopy: An Evaluation of Robustness for clinical Deep-Learning

David Kügler DAVID.KUEGLER@GRIS.TU-DARMSTADT.DE
Department of Computer Science, Technische Universität Darmstadt, Germany

Andreas Bucher ANDREASMICHAEL.BUCHER@KGU.DE
Department of Diagnostic and Interventional Radiology, University Hospital Frankfurt, Germany

Johannes Kleemann JOHANNES.KLEEMANN@KGU.DE
Department of Dermatology, Venereology and Allergology, University Hospital Frankfurt, Germany

Alexander Distergoft, Ali Jabhe, Marc Uecker, Salome Kazeminia, Johannes Fauser, Daniel Alte, Angeelina Rajkarnikar
Department of Computer Science, Technische Universität Darmstadt, Germany

Arjan Kuijper ARJAN.KUIJPER@IGD.FRAUNHOFER.DE
Fraunhofer IGD, Darmstadt, Germany

Tobias Weberschock TOBIAS.WEBERSCHOCK@KGU.DE
Markus Meissner MARKUS.MEISSNER@KGU.DE
Department of Dermatology, Venereology and Allergology, University Hospital Frankfurt, Germany

Thomas Vogl T.VOGL@EM.UNI-FRANKFURT.DE
Department of Diagnostic and Interventional Radiology, University Hospital Frankfurt, Germany

Anirban Mukhopadhyay ANIRBAN.MUKHOPADHYAY@GRIS.TU-DARMSTADT.DE
Department of Computer Science, Technische Universität Darmstadt, Germany

Editors: Under Review for MIDL 2019

Abstract

Deep Learning (DL)-based diagnostic systems are getting approved for usage as fully automatic or secondary opinion products. This development derives from the achievement of expert-level performance by DL across several applications (e.g. dermoscopy and diabetic retinopathy). While recent literature shows their vulnerability to imperceptible digital manipulation of the image data (e.g. through cyberattacks), the performance of medical DL systems under *physical world attacks* is not yet explored. This problem demands attention if we want to safely translate medical DL research into clinical practice. In this paper, we design the first small-scale prospective evaluation addressing the vulnerability of DL-dermoscopy systems under physical world attacks in absentia of knowledge about the underlying DL-architecture. We publish the entire dataset of collected images as Physical Attacks on Dermoscopy (PADv1) for public use. The evaluation of susceptibility and robustness reveals that such attacks lead to on average 31% accuracy loss across popular DL-architectures. The DL diagnosis is changed by the attack in one of two cases even without any knowledge of the DL method.

Keywords: Dermoscopy, Vulnerabilities of Deep Learning, Adversarial Examples, Physical World Attacks, Real Clinical Attacks, Skin Cancer

1. Introduction

While medical systems empowered by Deep Learning (DL) are getting approved for clinical procedures in hospitals and practices alike, only recently their vulnerability is taken into consideration. In these studies (Paschali et al., 2018; Finlayson et al., 2018; Taghanaki et al., 2018; Kügler et al., 2018), the authors attack medical DL systems digitally (e.g. through cyberattacks) using adversarial examples. They show that maliciously crafted perturbations of the image can alter the DL diagnosis with high success rates while maintaining high confidence. Although demonstrated for computer vision tasks (Sharif et al., 2016), so far no attention is given to vulnerabilities of DL systems under physical world attacks in realistic clinical settings. While cyberattacks corrupt the image data digitally, physical world attacks are constrained to changing the appearance of the region under consideration in the real world. This leads to the following question: Can physical world attacks from the clinical setting severely affect the performance of popular DL architectures?

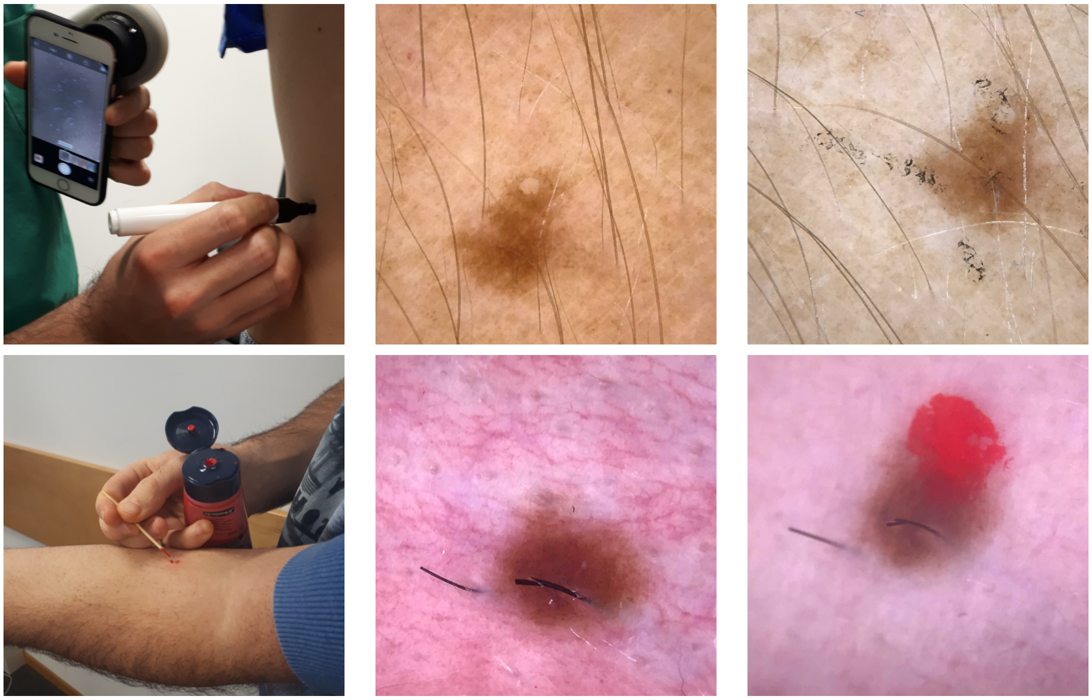
One prominent medical application of DL is dermoscopy, where the diagnostic performance reaches the level of experts (Esteva et al., 2017; Mahbod et al., 2018). Since the acquisition of dermoscopic images is quick and causes neither pain nor harm to test subjects, dermoscopy is an interesting test bench for prospective study of physical world attacks. In addition, dermoscopy is a likely candidate for fully automated[1] future DL-based products with the promise of significantly simplifying the analysis of skin lesions (Lowell et al., 2001). Emphasizing the relevance of physical world attacks, we identify exemplary clinical scenarios with severe consequences for health-care in the following: (1) to indicate additional testing incurring extra fees, (2) to mark the lesion with a little dot to help the system zero-in or just to help with locating skin lesions (see Figure 1), (3) to extort for personal gain or (4) to taint the reputation of medical software.

We propose a new evaluation for DL systems in medical imaging: *robustness to physical attacks*. Such a prospective evaluation is necessary for the translation of DL into clinical products, since seemingly irrelevant details from day-to-day clinical practice can lead to the failure of DL-systems, despite not being particularly compelling for learning theory research. Designing such experiments, which simulate physical world attacks in a clinical setting, requires to overcome a central challenge: the wealth of different physical corruptions. In dermoscopy, this includes natural and synthetic artifacts such as bruises, little dots and lines. In this paper, we systematically analyze *seven* different attack patterns by creating a small-scale prospective evaluation dataset (Physical Attacks on Dermoscopy, PADv1) containing both clean and physically attacked images[2]. We assume that the attacker has no knowledge of underlying DL architecture, model weights or training data. The evaluation demonstrates how nonspecific knowledge of the task can be leveraged to fool the decision method. Moreover, these attacks neither inhibit the dermatologist's ability of accurate diagnosis nor add visual characteristics specific to a distinct skin disease (e.g. malignant melanoma). We evaluate the performance of *five* popular DL architectures including ResNet, InceptionV3, InceptionResNetV2, Xception and MobileNet.

Wherever there is money to be made, some people will exploit the opportunity and abuse ambiguities, which is shown by cyber threats (Jarrett et al., 2018; American Medical

[1] https://www.reuters.com/article/us-fda-ai-approval/u-s-fda-approves-ai-device-idUSKBN1HI2LC

[2] Watch the video of our Physical Attacks on Dermoscopy: `https://youtu.be/eyA6flBjCfQ`



attack procedure

clean image (InceptionV3 successful)

attacked image (InceptionV3 failed)

Figure 1: Physical Attacks on Dermoscopy (PAD): left: creation for physical world artifacts in practice, center: clean, correctly classified images; right: corresponding attacked, incorrectly classified images

Association, 2017) as well as individual (Ornstein and Grochowski Jones, 2014) and institutional (Kalb, 1999) instances of fraud and abuse in healthcare. Since we are interested in making DL-based products clinically useful, more prospective clinical evaluations such as the one proposed here are required (Kaplan and Irvin, 2015). For the benefit of the medical imaging community, we make PADv1 publicly available[3] for further evaluation.

2. Related Work

A significant body of work (Goodfellow et al., 2014; Kurakin et al., 2018; Moosavi-Dezfooli et al., 2016; Brendel et al., 2018) presents methods for digitally attacking DL by adversarial examples. In the context of adversarial attacks, we differentiate between black-box and white-box attacks, which differ in the available knowledge about the attacked DL system. While we have no knowledge in the former, we assume complete knowledge in the latter. Digital attacks have successfully been applied to medical applications including dermoscopy (Paschali et al., 2018; Taghanaki et al., 2018; Finlayson et al., 2018; Kügler et al., 2018).

[3] Download the dataset at https://www.gris.tu-darmstadt.de/short/PhysicalAttacksonDermoscopy

Physical world attacks, however, are less common. The transfer to the physical world focuses on printed objects (Kurakin et al., 2016), which are misclassified when acquired by a camera. An attacker can even fool commercial DL systems for facial recognition by preparing a custom-made frame for glasses (Sharif et al., 2016). Both methods realize white-box attacks. Sharif et al. (2016) optimizes the texture printed onto the frames forward-facing side through the simulation of different head-glasses configurations. Eykholt et al. (2017) propose Robust Physical Perturbations (RP_2) for road sign recognition. They achieved high success rates for their sign classification task, even misclassifying one Stop Sign as a Speed Limit Sign under all testing conditions. Their work has been criticized by Lu et al. (2017), because in autonomous driving detectors are used instead of classifiers. More recently, the consistency to fool DL in varying conditions is improved to generate adversarials independent to a chosen distribution of transformations (Athalye et al., 2017).

However, to this day, no physical world black-box attack is reported in clinical setting.

3. Methods

We place small artifacts such as dots and lines in a region around the skin lesion and evaluate, whether the DL diagnosis changes.

3.1. Models and Training

As is common for dermoscopic classification tasks (Codella et al., 2017), we fine-tune multiple state-of-the-art neural network architectures pre-trained on the ImageNet dataset: ResNet (He et al., 2015), InceptionV3 (Szegedy et al., 2015), InceptionResNetV2 (Szegedy et al., 2016), Xception (Chollet, 2016) and MobileNet (Howard et al., 2017). The fine-tuned architecture consists of Dropout (75%), a fully-connected layer (1024 units, ReLU activation), Dropout (75%), and a fully-connected layer with 7 outputs and softmax activation behind an average-pooling layer. Models are trained for 85 epochs using cross-entropy loss and the RMSProp optimizer with the learning rate initially 10^{-3} and 10^{-5} by the final epoch. During the first 5 epochs, we keep the pre-trained parameters frozen. Additionally, we use class-weights to counteract the imbalanced class-distribution of the dataset.

For each architecture, we train 5 instances, resulting in a total of 25 trained networks in effect averaging out the stochastic properties of the training process. We use a 80/20 train/test-split of the dataset images and choose the network by highest validation score (balanced accuracy) in a keras environment with tensorflow backend.

Table 1: Details for PADv1-attacked: Number of images (unique lesion/attack combinations) per physical attack pattern

	Black		**Red**	
	Acrylic	**Pen**	**Acrylic**	**Pen**
Dot	17^F	13^C	17^F+21^C	14^C
Line	19^O	10^O		

^{F} Dot placed at distance to lesion
^{C} Dot placed in direct vicinity of lesion
^{O} Line partially overlapping lesion

3.2. Physical Attack Methodology

To simulate the physical attack in real clinical setting, we consider dots and lines of different color placed in relation to the lesion. While visually very different to specific skin diseases, the artifacts are chosen as proxies for shapes that correspond to specific observations. We develop three basic attack types: black dot, red dot and black line. The black dot has some resemblance with a nevus or melanoma, red dots could be misconstrued as blood stains or bruising. Finally, thicker, black lines look similar to hair. To uniquely identify different attack patterns, we use acronyms: BDFA, BDCP, BLOA, BLOP, RDFA, RDCA, RDCP. The first letter indicates the color (Black, Red), the second the shape (Dot, Line), the third the position (Close, Far, Overlapping) and the last indicates the texture (Acrylic, Pen). Table 1 shows all variations we use for our attacks.

The attack patterns are prepared after clean images are taken. To achieve different textures, artifacts are applied by different pens and acrylic paint. We draw and paint dots using two tools: a marker with a round tip of approx. 3 millimeter size (pen) or a sharp-tipped tool from wood (acrylic). We aim to keep it between half the size of the lesion and equivalent to size of the lesion (typically around 4 millimeters in diameter). As an additional degree of freedom for dots, we vary the relative position of the artifacts to the lesion. Dots are placed at the border of the lesion (close, direct vicinity) or at a distance of approximately the size of the lesion (far). For black lines, we use a black pen with 0.05 millimeter thickness (pen) and wooden splinters (acrylic). In some cases we use a sticker for easy identification and correspondence of clean and attacked images.

3.3. Datasets

To train our Deep Learning models, we use the publicly available training dataset HAM10000 (Tschandl et al., 2018; Codella et al., 2017). This unbalanced dataset contains 10015 dermoscopic lesion images. Experts classified individual images as one of seven skin diseases ranging from melanocytic nevi to vasular skin lesions; 53.3% of them are pathologically verified.

We introduce a new dataset: Physical Attacks on Dermoscopy (PADv1), which contains prospective clinical data of 51 clean images with verified ground truth (X_{clean}, PADv1-clean), 111 attacked images (X_{attacked}, PADv1-attacked, detailed in table 1) and a correspondence table linking clean and attacked images. These attacked images are generated based on the method outlined in section 3.2. PADv1 consists of data from young adults aged between 25 to 35 years, including a total of 8 participants (6 male and 2 female) from caucasian and middle eastern skin types. We use PADv1-clean i) to verify the generalization ability of the models in a prospective study and ii) as a reference for physical world attacks performed on the same lesions. PADv1-attacked contains corresponding images with lesion and attacking artifact for the analysis of physical world attacks.

PADv1 images are acquired and diagnosed by a dermatologist using a combination of a DermLite 2 Pro HR dermoscope and an iPhone 7. All lesions are non suspicious for melanoma. PADv1 is a small-scale dataset (a total of 162 images) not histologically verified and like the HAM10000 dataset (Tschandl et al., 2018) dominated by melanocytic nevi.

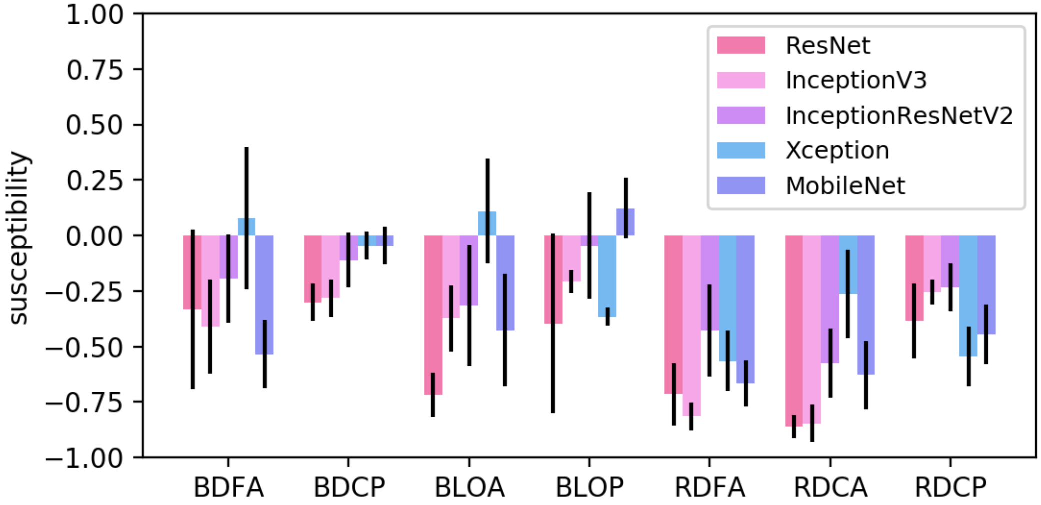


Figure 2: Susceptibility to physical world attacks, minimum: -1 (worst case, no attacked image correctly predicted); larger than 0: accuracy on PADv1-attacked larger than on PADv1-clean; BDFA: Black Dot Far Acrylic, BDCP: Black Dot Close Pen, BLOA: Black Line Over Acrylic, BLOP: Black Line Over Pen, RDFA: Red Dot Far Acrylic, RDCA: Red Dot Close Acrylic, RDCP: Red Dot Close Pen

4. Results

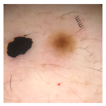
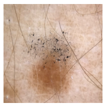
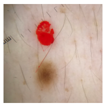
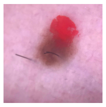
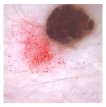
We test the performance of five popular architectures (ResNet, InceptionV3, Inception-ResNetV2, Xception, MobileNet) each from five independent trainings. We report the performance of five models to show that the effects are not statistical outliers to a specific set of weights, but rather agree across different training instances and architectures.

4.1. Quantitative analysis of physical attacks in PADv1

For the evaluation of the vulnerability to physical attacks, we introduce two evaluation metrics: susceptibility and robustness. Susceptibility compares the accuracy of the DL system under attack $\mathrm{acc}(X_{\mathrm{attacked}})$ with accuracy on clean data $\mathrm{acc}(X_{\mathrm{clean}})$: $\frac{\mathrm{acc}(X_{\mathrm{attacked}})}{\mathrm{acc}(X_{\mathrm{clean}})} - 1$. Robustness indicates the stability to a decision, i.e. the fraction of cases, where the decision is the same for the clean and the attacked image.

In terms of susceptibility, all models are negatively affected by attacked images with accuracies decreasing by up to 60% for some attacks. Although most architectures are affected negatively by the attack patterns, fig. 2 also indicates Xception and MobileNet less susceptible with values close to zero. The average *loss of accuracy* due to attacks across architectures is 30.8%. The figure shows standard deviations in addition to average susceptibilities.

Table 2: Robustness of different architectures to seven Physical World Attacks; Acronyms of attacks are listed in the caption of fig. 2; InceptRNV2: InceptionResNetV2

Arch\Attack	 **BDFA**	 **BDCP**	 **BLOA**	 **BLOP**	 **RDFA**	 **RDCA**	 **RDCP**	 **Overall**
ResNet	0.48	0.66	0.39	0.37	0.29	0.18	0.51	**0.40**
InceptionV3	0.54	0.63	0.46	0.63	0.21	0.14	0.64	**0.43**
InceptRNV2	0.72	0.74	0.65	0.60	0.48	0.36	0.70	**0.60**
Xception	0.69	0.95	0.65	0.63	0.31	0.42	0.44	**0.57**
MobileNet	0.40	0.86	0.56	0.60	0.28	0.34	0.53	**0.48**
Mean	**0.57**	**0.77**	**0.54**	**0.57**	**0.31**	**0.29**	**0.56**	**0.50**

With robustness $\sum_i N^{-1} \cdot 1(f(x_{i,\text{clean}}) = f(x_{i,\text{attacked}}))$, results (table 2) show that attacks change predictions in 40% to 60% of cases with two attacks (RDCA, RDFA) being effective in 70% of cases across different models. Comparing averages for this 111-image dataset, ResNet is the least robust, while InceptionResNetV2 and Xception are most robust.

We also report confidence values for the evaluation of PADv1-clean and PADv1-attacked in Figure 3. Unlike for digital adversarial examples, physical attacks present a difference of 5% in average confidence.

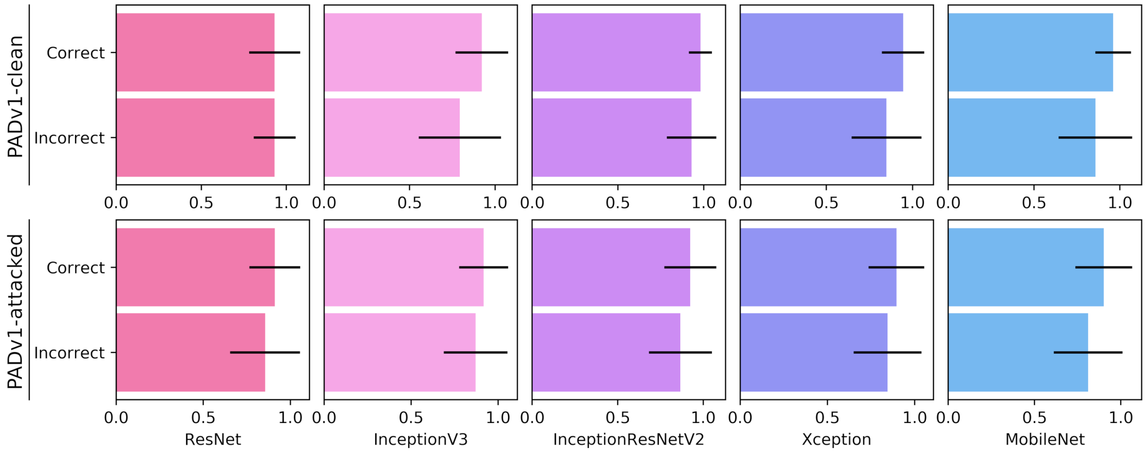


Figure 3: Mean and Standard deviation of network confidence based on its decision

Table 3: Mean and standard deviation of model accuracy on our 20% test-split from the HAM10000 dataset and the PADv1-clean dataset; five models per architecture

Architecture\DataSet	**HAM10000 (test split)**	**PADv1-clean**
ResNet	0.776 ± 0.0074	0.752 ± 0.1591
InceptionV3	0.753 ± 0.0114	0.758 ± 0.0685
InceptionResNetV2	0.778 ± 0.0062	0.876 ± 0.0423
Xception	0.758 ± 0.0107	0.830 ± 0.0837
MobileNet	0.765 ± 0.0034	0.906 ± 0.0357

4.2. Performance on clean data

To emphasize the quality of the proposed models in clean dermoscopy cases, we report therin performance on both the HAM10000 (Tschandl et al., 2018) and the PADv1-clean dataset. Since seven classes exist for dermoscopy, we report the balanced accuracy on the HAM10000 dataset evaluation. However, since PADv1 only includes a small subset of lesion types, the evaluation of “balanced accuracy” is impossible. Table 3 therefore includes the balanced accuracy for HAM10000 and accuracy for PADv1-clean. Considering we do not introduce sophisticated data augmentation or ensembling to our models, they perform comparable to state-of-the-art methods, such as presented by Codella et al. (2017).

5. Discussion

In this paper, we present a novel evaluation (including a new dataset PADv1 and two metrics) for the robustness of DL methods in clinical settings. We show small artifacts captured from the real world can significantly reduce the accuracy of DL diagnosis where dermatologists would not be impacted.

While limited in scope (dataset small), the success of our physical attacks unveils the potential for exploitation of the healthcare system by corrupt or malicious patients and medical staff. If unexplored, even trained experts would remain unaware of such threats, because in addition to malicious scenarios, consequences might arise accidentally for random or patient-specific scenarios: These might include parts of a tattoo, a bruise, a scar, a scab, improper preparation or even a patch covering these up.

Understanding why attacks are successful – independent to whether they are digital or physical – is crucial to make medical decision processes safe. In future work, we will focus on robustness of methods in medical image analysis: To this effect, we will explore robust methods (ensembling, confidence evaluations), the underlying reasoning (region-dependent prediction, interpretability and causal models) and a general understanding of mechanism in adversarial examples (Smith and Gal, 2018; Kügler et al., 2018).

References

American Medical Association. 8 in 10 doctors have experienced a cyberattack in practice, 2017. URL https://www.ama-assn.org/practice-management/sustainability/8-10-doctors-have-experienced-cyberattack-practice.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples, 2017. URL https://arxiv.org/pdf/1707.07397.

Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliqi, Marcel Salathé, Sharada P. Mohanty, and Matthias Bethge. Adversarial vision challenge, 2018. URL https://arxiv.org/pdf/1808.01976.

François Chollet. Xception: Deep learning with depthwise separable convolutions, 2016. URL https://arxiv.org/pdf/1610.02357.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), 2017. URL https://arxiv.org/pdf/1710.05006.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models, 2017. URL http://arxiv.org/pdf/1707.08945.

Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane, and Andrew L. Beam. Adversarial attacks against medical deep learning systems, 2018. URL http://arxiv.org/pdf/1804.05296.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014. URL http://arxiv.org/pdf/1412.6572.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/pdf/1512.03385.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. URL https://arxiv.org/pdf/1704.04861.

Mark Jarrett, Navid Ghaffarzadegan, Mohammad S. Jalali, and Jessica P. Kaiser. Cybersecurity in hospitals: A systematic, organizational perspective. *Journal of Medical Internet Research*, 20(5), 2018.

Paul E. Kalb. Health care fraud and abuse. *JAMA*, 282(12):1163, 1999.

Robert M. Kaplan and Veronica L. Irvin. Likelihood of null effects of large nhlbi clinical trials has increased over time. *PloS one*, 10(8):e0132382, 2015.

David Kügler, Alexander Distergoft, Arjan Kuijper, and Anirban Mukhopadhyay. Exploring adversarial examples: Patterns of one-pixel attacks. In Danail Stoyanov, Zeike Taylor, Seyed Mostafa Kia, Ipek Oguz, Mauricio Reyes, Anne Martel, Lena Maier-Hein, Andre F. Marquand, Edouard Duchesnay, Tommy Löfstedt, Bennett Landman, M. Jorge Cardoso, Carlos A. Silva, Sergio Pereira, and Raphael Meier, editors, *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, volume 11038 of *Lecture Notes in Computer Science*, pages 70–78. Springer International Publishing, Cham, 2018. ISBN 978-3-030-02627-1.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2016. URL https://arxiv.org/pdf/1607.02533.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan Yuille, Sangxia Huang, Yao Zhao, Yuzhe Zhao, Zhonglin Han, Junjiajia Long, Yerkebulan Berdibekov, Takuya Akiba, Seiya Tokui, and Motoki Abe. Adversarial attacks and defences competition, 2018. URL https://arxiv.org/pdf/1804.00097.

B. A. Lowell, C. W. Froelich, D. G. Federman, and R. S. Kirsner. Dermatology in primary care: Prevalence and patient disposition. *Journal of the American Academy of Dermatology*, 45(2):250–255, 2001.

Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles, 2017. URL https://arxiv.org/pdf/1707.03501.

Amirreza Mahbod, Gerald Schaefer, Isabella Ellinger, Rupert Ecker, Alain Pitiot, and Chunliang Wang. Fusing fine-tuned deep features for skin lesion classification. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 71:19–29, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR 2016), IEEE*, Piscataway, NJ, 2016. IEEE. ISBN 1467388521.

Charles Ornstein and Ryann Grochowski Jones. Top billing: Meet the docs who charge medicare top dollar for office visits, 2014. URL https://www.propublica.org/article/billing-to-the-max-docs-charge-medicare-top-rate-for-office-visits.

Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples. In Alejandro Frangi, Julia Schnabel, Christos Davatsikos, Carlos Alberola-López, Gabor Fichtinger, Alejandro F. Frangi, Julia A. Schnabel, and Christos Davatzikos, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, Image Processing, Computer Vision, Pattern Recognition, and Graphics, pages 493–501. Springer International Publishing, Cham, 2018. ISBN 978-3-030-00937-3.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime. In Stefan Katzenbeisser and Edgar Weippl, editors, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540, New York, NY, 2016. Association for Computing Machinery. ISBN 9781450341394.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection, 2018. URL http://arxiv.org/pdf/1803.08533.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. URL https://arxiv.org/pdf/1512.00567.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016. URL https://arxiv.org/pdf/1602.07261.

Saeid Asgari Taghanaki, Arkadeep Das, and Ghassan Hamarneh. Vulnerability analysis of chest x-ray image classification against adversarial attacks, 2018. URL http://arxiv.org/pdf/1807.02905.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.