

# Model Fooling Attacks Against Medical Imaging: A Short Survey

***Tuomo Sipola*** (✉), ***Samir Puuska***, ***Tero Kokkonen***

*JAMK University of Applied Sciences, Jyväskylä, Finland,  
<https://www.jamk.fi/en/Home/>*

## ABSTRACT:

This study aims to find a list of methods to fool artificial neural networks used in medical imaging. We collected a short list of publications related to machine learning model fooling to see if these methods have been used in the medical imaging domain. Specifically, we focused our interest to pathological whole slide images used to study human tissues. While useful, machine learning models such as deep neural networks can be fooled by quite simple attacks involving purposefully engineered images. Such attacks pose a threat to many domains, including the one we focused on since there have been some studies describing such threats.

## ARTICLE INFO:

RECEIVED: 08 JUN 2020

REVISED: 31 AUG 2020

ONLINE: 13 SEP 2020

## KEYWORDS:

deep learning, artificial neural networks,  
adversarial images, machine learning,  
medical imaging



Creative Commons BY-NC 4.0

## Introduction

Artificial Intelligence (AI) based solutions, especially deep learning based on neural networks, are widely used in the medical domain. For example, AI is used for helping and automatizing cancer diagnosis based on image data. The benefit of this approach is to relieve experts to work on more important tasks while automated systems can inspect images and give initial recommendations.

If an attacker can fool the AI processing, ramifications can be devastating. Such attacks may result in incorrect treatment procedures, causing extreme

circumstances with a worst-case scenario of losing human lives. In addition, wrong diagnoses could undermine the public trust in medical professionals. This paper presents a short survey of model fooling attacks against neural networks in the medical domain.

Fooling neural networks is an important subject because machine learning models are widely used in medicine for automating many processes and for helping with diagnosis. For example, Rai et al. proposed a convolutional neural network for healthcare assistant<sup>29</sup> while Rastgar-Jazi and Fernando used neural networks for detecting heart abnormalities from electrocardiogram (ECG) data.<sup>30</sup> Similarly, authors of<sup>31</sup> used neural networks for prediction and prevention of heart attacks from ECG data while Murugesan and Sukanesh used neural networks for detecting brain tumours in electroencephalograms (EEG) signals.<sup>24</sup> Syam and Marapareddy discussed three different scenarios of classification problems, where one is skin lesion (cancer) classification from images.<sup>34</sup> As can be seen, these machine learning solutions are useful for many medical applications.

Effectiveness of neural network based deep learning is based on the used algorithm and learning data. If learning dataset is inadequate or contains incorrect information, results will be inaccurate. Similarly, if there are known weaknesses in the used algorithm, they can be compromised. In that sense, AI components can be attacked and fooled to behave incorrectly. As an example of a weakness, Afifi and Brown explore how white balance of photography impact the performance of deep neural networks,<sup>2</sup> while authors of<sup>21</sup> generated adversarial noise for fooling the neural networks. Gu et al. discussed about gradient shielding method for understanding the vulnerabilities in neural networks.<sup>13</sup> MOEA-APGA is an algorithm for achieving targeted attacks against neural networks,<sup>9</sup> and another similar algorithm is called DeepFool implemented for computing perturbations that fool neural networks.<sup>23</sup> In a medical domain, Chuquicusma et.al. studied about fooling radiologists for lung cancer diagnosis.<sup>8</sup> As can be seen, many such attack vectors exist.

As a powerful machine learning method, deep learning has also been applied to images related to pathology, for example, trying to classify images of cancer whole slide images (WSI). Serag et al. present an overview of the application of artificial intelligence for pathology and tissue analytics.<sup>32</sup> As another example, convolutional neural networks have been used for nuclear segmentation, which is an important part of tissue cancer grading.<sup>18</sup> Deliberately produced wrong segmentation could result in wrong diagnoses. Pre-trained convolutional neural networks have been compared to training from scratch using the Kimia Path24 dataset, with results indicating that pre-trained networks are quite competitive.<sup>15</sup> Using such pre-trained models creates a possibility of hidden attacks trained into the model or abusing known deficiencies of such models.

In this study, we collected a list of relevant research papers concerning medical imaging and attacks against neural networks. We queried the publicly available Google Scholar database to identify publications relevant to deep neural network fooling, deep neural networks in medical imaging and deep neural

networks fooled in that domain. The results of this short survey should be useful for anyone trying to understand the vulnerabilities of neural networks in specific domains. Moreover, the use of them in medical imaging raises the question of reliability and robustness when targeted by such attacks. As can be seen, targeted attack against neural networks in medical domain is a realistic scenario. From the attacker perspective, medical domain can be considered as a valuable target because of the critical ramifications of possible attack. In addition, there are known vulnerabilities with neural networks that are highly used in medical domain.

Below, we present the medical imaging domain and then discuss about machine learning regarding that domain. Next sections describe the state of fooling deep neural networks and how it has been applied to the medical domain. Finally, we present our concluding thoughts.

### **Short Introduction to Whole Slide Images in Cancer Diagnosis**

Quick and affordable laboratory cancer diagnosis methods are of great importance. One of the well-established methods is light microscopy with a stain, such as haematoxylin and eosin (H&E). The H&E stain makes various tissue components visible, allowing diagnosis based on e.g. their morphological features.<sup>14</sup>

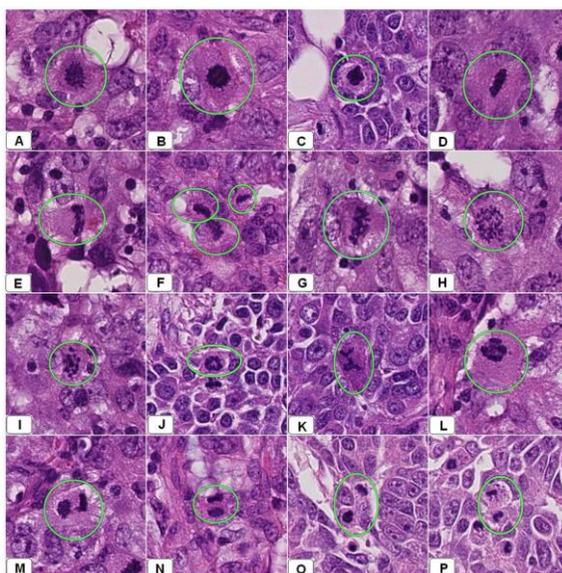
The advent of digital pathology and whole-slide imaging (WSI) have provided a computerized way to analyse and share the results of light microscopy. By digitizing the tissue images, a variety of automated methods can be used to perform image analysis, annotation, and workflow improvements. Turning glass slides to a digital format requires a slide scanner, which digitizes the slide using specialized format that allows e.g. various zoom levels and metadata to be stored in one data file. This data can then be easily shared, further processed using a variety of tools, and even easily used in teaching in a virtualized microscopy environment.<sup>1</sup>

Distinguishing between benign and malignant tumours is essential for accurate prognosis. One of the features that separate the two is differentiation and anaplasia. In general, benign tumours consist of cells that resemble the tissue where they originated from. They retain much of the functionality and morphology of their non-transformed counterparts but may invade surrounding tissue. Malignant tumours, on the other hand, lose their resemblance to their normal counterparts and become undifferentiated (anaplastic). This change results in noticeable change in cell morphology, and it is possible to observe this using light microscopy and stains. These observable changes include variations in size and shape, nuclear abnormalities and atypical mitoses. Assigning a value to this differentiation is called grading. The criteria and schemes are dependent on the type of tumor.<sup>19</sup>

For breast cancer, observing mitoses has been shown to be a good predictor for tumour development and prognosis. In order to proliferate, tumour cells need to overcome various limitations that prevent ordinary cells from dividing indefinitely. These mutations may result in increased cell-cycle activity, and even majorly affect the mitosis process itself by causing atypical-looking cell

divisions which may be visually observed using light microscopy.<sup>27</sup> Figure 1 shows an example of breast cancer WSI from Al-Janabi et al.<sup>4</sup>

Detecting abnormal morphology and quantifying the number of various cell features is a good candidate for automatization via machine learning and computer vision methods. WSI with sufficient quality can be automatically annotated. Digital pathology is expected to improve convenience and quality of the process. Nam et al. provide an introduction to digital pathology aimed at healthcare professionals.<sup>25</sup> Furthermore, Komura and Ishikawa made a short review of machine learning methods for histopathological image analysis, listing seven whole slide image datasets and 21 hand annotated histopathological datasets.<sup>16</sup>



**Figure 1: Example of WSI showing several breast resections with infiltrative ductal carcinoma.**

Figure courtesy of Al-Janabi et al.,<sup>4</sup> distributed under the terms of the Creative Commons Attribution License.

## Fooling Deep Neural Networks

Deep neural networks and deep learning in general refers to a field of study, where complex concepts are learned from simpler representations by creating an interconnected network of activation functions and weights.<sup>11</sup> Due to their nature, these networks may contain flaws which make them susceptible to various classes of errors. These imperfections may be used maliciously to force the network into making an erroneous prediction. There have been several successful attempts at creating methods to fool deep neural networks.

One approach is to give an adversarial image as input to the classifier. Nguyen, Yosinski and Clune used an evolutionary algorithm optimization method to generate unrecognizable images to the human eye. Those images fooled a neural network to classify them as an object with high certainty, even though it should not have. They describe these images as costly exploits that could be used against deep neural networks.<sup>26</sup> Moosavi-Dezfooli, Fawzi and Frossard propose the DeepFool algorithm that efficiently generates adversarial images and quantifies the robustness of image classifiers.<sup>23</sup> An adversarial image is wrongly classified as something else than what the image clearly represents to the human eye. In the paper, a slight perturbation was added to an image of an animal to misclassify a whale as a turtle. Furthermore, it is possible to create adversarial 2D images robust to noise, distortion and affine transformations, and even adversarial 3D printed objects (a turtle).<sup>6</sup>

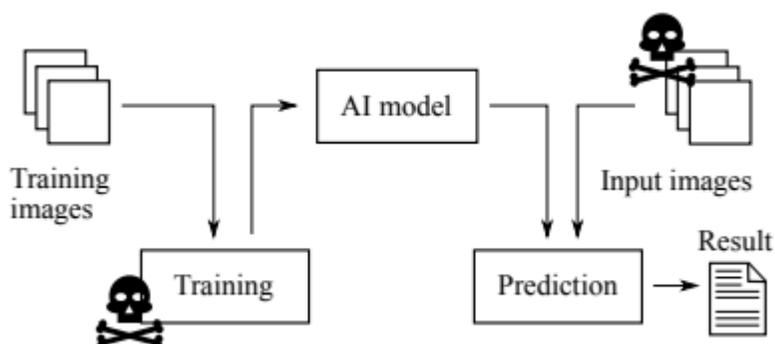
Adversarial patches are images that can be placed inside another image to fool a neural network classifier. Brown et al. have shown the effectiveness of such images.<sup>7</sup> It is easy to see that inserting such patches to medical images could yield similar results, resulting in a false classification.

Research has already addressed cases of changing only one pixel of an image to cause it to be classified as another object.<sup>33</sup> It is remarkable that a change of colour in one pixel could fool the neural network. A move towards a more theoretical understanding of one pixel-attacks and incorrect mapping to low dimensional manifold has also been proposed. This makes it easy to find localized areas where one-pixel attacks should be more effective.<sup>17</sup>

Backdoored images can be created when attacking the learning stage of a neural network. These malign models can be deployed to production, and the fault is only revealed when the bad image is given as an input, resulting in wrong classification. Outsourced training opens the possibility of creating backdoored neural networks that behave badly on input specified by the attacker.<sup>12</sup> In a similar scheme, called poisoning attack, artificially poisoned data being sent to a model gradually change the model to conform to the attacker's goals. Yang et al. used an autoencoder (instead of the more traditional direct gradient method) to generate poisoned input data for deep neural networks.<sup>37</sup>

The evident vulnerability of neural networks against several types of attacks is alarming because these methods are being proposed in several real-world domains. See <sup>3</sup> for a survey of adversarial attacks against deep learning in computer vision. The authors not only list several attacks but also include defences. They conclude that there is a threat against safety and security critical applications.

Figure 2 shows a schematic presentation of the possible attack routes described above. In this paper we have identified two parts of the AI process, which could be targeted.



**Figure 2: A schematic picture of AI model fooling.**

The input images could be altered using adversarial images, patches or one-pixel attacks. In addition, the training process itself could be tampered with.

### Fooling Deep Learning in Medical Imaging and Pathology

Although a rather new concern, the vulnerabilities intrinsic to neural network solutions have been identified by the medical community. For evident health reasons the accuracy and robustness of methodology in the medical domain is very important. Tizhoosh and Pantanowitz list challenges and opportunities related to artificial intelligence and digital pathology. One of the challenges concerns adversarial attacks and the shakiness of deep decisions made by neural networks.<sup>36</sup> This fundamental lack of robustness could be one avenue of future research.

Adversarial examples in medical imaging can change the behaviour of classifiers and segmentation, illustrating the lack of robustness in the neural network models. Such approach can also be used for model evaluation.<sup>28</sup> Vulnerability during segmentation could lead to wrong representation of reality during the following stages of diagnosis. Again, the less understood and erratic boundaries of classification are a concern that enable an attack vector.

Deep learning networks classifying X-ray images are also vulnerable to attacks.<sup>35</sup> Being perhaps the most familiar scenario to the public, X-ray image processing is a natural target for automation. However, these kinds of perturbation attacks show that the models can be fooled.

Finlayson et al. successfully use adversarial attacks against medical imaging in three domains: funduscopy, chest X-ray images and dermoscopy. They also present a risk model for the machine learning pipeline.<sup>10</sup> Patch attacks and projected gradient descent both seem to work against real world images, reducing the reliability of the classifier. However, neural networks can be made robust against perturbation attacks by exploiting the structure of the optimization task.<sup>20</sup>

Not all uses of these methods are harmful. It is possible to use existing medical imaging data to generate more training data and to tackle uneven class balance using various methods, including generative adversarial networks.<sup>22</sup> The methods described above can also be used for beneficial inpainting of missing areas in biomedical imaging. Armanious et al. used generative adversarial networks to inpaint missing areas or incomplete medical images.<sup>5</sup>

The identified fooling methods are listed in Table 1. As can be seen, some of the fooling methods have been used in the medical domain. It should be noted that training process tampering is probably more difficult to execute in practice.

**Table 1. Fooling methods against deep neural networks and those in the medical domain.**

Method	References	Medical domain
Adversarial images	[26],[23],[6]	[28],[35],[10],[22],[5]
Adversarial patches	[7]	[10]
One-pixel attack	[33],[17]	
Training process tampering	[12],[37]	

## Conclusions

Although modern neural networks have proven useful for detecting cancerous cell growth, it is possible to mislead these algorithms. There have been research exploits against deep learning methods, even in the field of pathology. Such exploits include specifically engineered adversarial images, adversarial patches put on actual images, one-pixel attacks and attacks focusing on fooling the training process. The scientific studies this short survey inspected include all those attacks. Even medical imaging is not safe from them, which promotes further study of the underlying causes and robustness problems stemming from the structure of neural networks. The expert opinions from the medical community will also broaden the understanding of the effect of these types of attacks.

## Acknowledgements

This research is partially funded by the Cyber Security Network of Competence Centres for Europe (CyberSec4Europe) project of the Horizon 2020 SU-ICT-03-2018 program.

## References

- 1 Famke Aeffner, Mark Zarella, Nathan Buchbinder, Marilyn M Bui, Matthew Goodman, Douglas Hartman, Giovanni Lujan, Mariam Molani, Anil Parwani, Kate Lillard, Oliver Turner, Venkata Vemur, Ana Yuil-Valdes, and Douglas Bowman, "Introduction to Digital Image Analysis in Whole-slide Imaging: A White Paper from the Digital Pathology Association," *Journal of pathology informatics* 10, no. 1 (2019): 9.

- <sup>2</sup> Mahmoud Afifi and Michael S Brown, "What Else Can Fool Deep Learning? Addressing Color Constancy Errors on Deep Neural Network Performance," In: *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 243–252.
- <sup>3</sup> Naveed Akhtar and Ajmal Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," *IEEE Access* 6 (2018): 14410–14430.
- <sup>4</sup> Shaimaa Al-Janabi, Henk-Jan van Slooten, Mike Visser, Tjeerd van der Ploeg, Paul J. van Diest, and Mehdi Jiwa, "Evaluation of Mitotic Activity Index in Breast Cancer Using Whole Slide Digital Images," *PloS one* 8, no. 12 (2013).
- <sup>5</sup> Karim Armanious, Youssef Mecky, Sergios Gatidis, and Bin Yang, "Adversarial Inpainting of Medical Image Modalities," In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 3267–3271.
- <sup>6</sup> Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, "Synthesizing Robust Adversarial Examples," In: *Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholm, Sweden, vol. 80*, 2018, pp. 284–293.
- <sup>7</sup> Tom B. Brown, Dandelion Mané, Aurko Roy, Martin Abadi, and Justin Gilmer, "Adversarial Patch," arXiv e-prints (2018), <https://arxiv.org/abs/1712.09665>.
- <sup>8</sup> Maria Chuquicusma, Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci, "How to Fool Radiologists with Generative Adversarial Networks? A Visual Turing Test for Lung Cancer Diagnosis," In: *2018 IEEE 15<sup>th</sup> international symposium on biomedical imaging (ISBI 2018)*, IEEE, 2018, pp. 240–244.
- <sup>9</sup> Yepeng Deng, Chunkai Zhang, and Xuan Wang, "A Multi-objective Examples Generation Approach to Fool the Deep Neural Networks in the Black-box Scenario," In: *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, IEEE, 2019, pp. 92–99.
- <sup>10</sup> Samuel Finlayson, Hyung Won Chung, Isaac Kohane, and Andrew Beam, "Adversarial Attacks Against Medical Deep Learning Systems," *arXiv e-print* (4 Feb. 2019), <https://arxiv.org/abs/1804.05296>.
- <sup>11</sup> Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (MIT Press, 2016), <http://www.deeplearningbook.org>.
- <sup>12</sup> Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg, "Badnets: Evaluating Backdooring Attacks on Deep Neural Networks," *IEEE Access* 7 (2019): 47230–47244.
- <sup>13</sup> Zhaoquan Gu, Weixiong Hu, Chuanjing Zhang, Hui Lu, Lihua Yin, and Le Wang, "Gradient Shielding: Towards Understanding Vulnerability of Deep Neural Networks," *IEEE Transactions on Network Science and Engineering*, 2020.
- <sup>14</sup> Luis Carlos Junqueira and Jose Carneiro, *Basic Histology Text & Atlas* (McGraw-Hill Professional, 2005).
- <sup>15</sup> Brady Kieffer, Morteza Babaie, Shivam Kalra, and H.R. Tizhoosh, "Convolutional Neural Networks for Histopathology Image Classification: Training vs. Using Pre-trained Networks," In: *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, IEEE, 2017, pp. 1–6.
- <sup>16</sup> Daisuke Komura, Shumpei Ishikawa, "Machine Learning Methods for Histopathological Image Analysis," *Computational and structural biotechnology journal* 16 (2018): 34–42.

- 17 David Kügler, Alexander Distergoft, Arjan Kuijper, and Anirban Mukhopadhyay, "Exploring Adversarial Examples," In: *Understanding and Interpreting Machine Learning in Medical Image Computing Applications* (Springer, 2018), 70–78.
- 18 Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi "A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology," *IEEE Transactions on Medical Imaging* 36, no. 7 (2017): 1550–1560.
- 19 Vinay Kumar, Abul Abbas, and Jon Aster, *Robbins Basic Pathology*, 10th edition (Saunders, Philadelphia: Elsevier, 2017).
- 20 Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv e-prints*, arXiv:1706.06083 (2019).
- 21 Marko Mihajlović and Nikola Popović, "Fooling a Neural Network with Common Adversarial Noise," In: *2018 19th IEEE Mediterranean Electrotechnical Conference (MELECON)*, IEEE, 2018, pp. 293–296.
- 22 Agnieszka Mikołajczyk and Michał Grochowski, "Data Augmentation for Improving Deep Learning in Image Classification Problem," In: *2018 international interdisciplinary PhD workshop (IIPhDW)*, IEEE, 2018, pp. 117–122.
- 23 Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, "Deepfool: A Simple and Accurate Method to Fool Deep Neural Networks," In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2016, pp. 2574–2582.
- 24 Muthukumar Murugesan and R. Sukanesh, "Automated Detection of Brain Tumor in EEG Signals Using Artificial Neural Networks," In: *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*, IEEE, 2009, pp. 284–288.
- 25 Soojeong Nam, Yosep Chong, Chan Kwon Jung, Tae-Yeong Kwak, Ji Youl Lee, Jihwan Park, Mi Jung Rho, and Heounjeong Go, "Introduction to Digital Pathology and Computer-aided Pathology," *The Korean Journal of Pathology* 54, no. 2 (2020): 125–134.
- 26 Anh Nguyen, Jason Yosinski, and Jeff Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2015, pp. 427–436.
- 27 Ryuji Ohashi, Shigeki Namimatsu, Takashi Sakatani, Zenya Naito, Hiroyuki Takei, and Akira Shimizu, "Prognostic Utility of Atypical Mitoses in Patients with Breast Cancer: A Comparative Study with ki67 and Phosphohistone h3," *Journal of surgical oncology* 118, no. 3 (2018): 557–567.
- 28 Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab, "Generalizability vs. Robustness: Adversarial Examples for Medical Imaging," *arXiv e-prints* (2018).
- 29 Siddhant Rai, Akshayanand Raut, Akash Savaliya, and Radha Shankarmani, "Darwin: Convolutional Neural Network based Intelligent Health Assistant," In: *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, 2018, pp. 1367–1371.

- <sup>30</sup> Maryam Rastgar-Jazi and Xavier Fernando, "Detection of Heart Abnormalities via Artificial Neural Network: An Application of Self Learning Algorithms," In: *2017 IEEE Canada International Humanitarian Technology Conference (IHTC), IEEE, 2017*, pp. 66–69.
- <sup>31</sup> Deepika Ravish, Nayana R Shenoy, K. Shanthi, and S. Nisargh, "Heart Function Monitoring, Prediction and Prevention of Heart Attacks: Using Artificial Neural Networks," In: *2014 International Conference on Contemporary Computing and Informatics (IC3I), IEEE, 2014*, pp. 1–6.
- <sup>32</sup> Ahmed Serag, Adrian Ion-Margineanu, Hammad Qureshi, Ryan McMillan, Marie-Judith Saint Martin, Jim Diamond, Paul O'Reilly, and Peter Hamilton, "Translational AI and Deep Learning in Diagnostic Pathology," *Frontiers in Medicine* 6 (2019): 185.
- <sup>33</sup> Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi, "One Pixel Attack for Fooling Deep Neural Networks," *IEEE Transactions on Evolutionary Computation* 23, no. 5 (2019): 828–841.
- <sup>34</sup> Rajarshi Syam and R. Marapareddy, "Application of Deep Neural Networks in the Field of Information Security and Healthcare," In: *2019 SoutheastCon, IEEE, 2019*, pp. 1–5.
- <sup>35</sup> Saeid Asgari Taghanaki, Arkadeep Das, and Ghassan Hamarneh. "Vulnerability Analysis of Chest X-ray Image Classification against Adversarial Attacks," In: *Understanding and Interpreting Machine Learning in Medical Image Computing Applications* (Springer, 2018): 87–94.
- <sup>36</sup> Hamid Reza Tizhoosh and Liron Pantanowitz, "Artificial Intelligence and Digital Pathology: Challenges and Opportunities," *Journal of Pathology Informatics* 9 (2018).
- <sup>37</sup> Chaofei Yang, Qing Wu, Hai Li, and Yiran Chen, "Generative Poisoning Attack Method against Neural Networks," *arXiv e-prints* (2017).