One-Pixel Attack Deceives Computer-Assisted Diagnosis of Cancer

JONI KORPIHALKOLA, Institute of Information Technology, JAMK University of Applied Sciences, Finland

TUOMO SIPOLA, Institute of Information Technology, JAMK University of Applied Sciences, Finland

SAMIR PUUSKA, Faculty of Information Technology, University of Jyväskylä, Finland

TERO KOKKONEN, Institute of Information Technology, JAMK University of Applied Sciences, Finland

Computer vision and machine learning can be used to automate various tasks in cancer diagnostic and detection. If an attacker can manipulate the automated processing, the results can be devastating and in the worst case lead to wrong diagnosis and treatment. In this research, the goal is to demonstrate the use of one-pixel attacks in a real-life scenario with a real pathology dataset, TUPAC16, which consists of digitized whole-slide images. We attack against the IBM CODAIT's MAX breast cancer detector using adversarial images. These adversarial examples are found using differential evolution to perform the one-pixel modification to the images in the dataset. The results indicate that a minor one-pixel modification of a whole slide image under analysis can affect the diagnosis by reversing the automatic diagnosis result. The attack poses a threat from the cyber security perspective: the one-pixel method can be used as an attack vector by a motivated attacker.

CCS Concepts: • **Applied computing** → **Imaging**; • **Security and privacy**; • **Computing methodologies** → *Supervised learning by classification*; • **Mathematics of computing** → *Evolutionary algorithms*;

Additional Key Words and Phrases: adversarial examples, cyber security, machine learning, medical imaging, breast cancer, model safety

1 INTRODUCTION

1.1 Cancer detection as a target

Cancer is one of the most common causes of death in the western world. The number of detected cancers of a determined type in a defined population during a year is expressed as cancer incidence rate (CIR), commonly formed as the number of cancers per 100,000 population [22]. According to the U.S. National Cancer Institute at the National Institutes of Health, the CIR, based on 2013–2017 statistics in the U.S., is 442.4 per 100,000 men and women per year [23]. In many common types of cancers, early detection is a key factor in improving the prognosis [24]. However, early detection is a time-consuming activity. Automating some of the work with techniques such as machine learning and computer vision can lead to faster detection, increased throughput, and reduced costs [10, 29], e.g., when a count of cells needs to be made from an image [26]. Although various approaches for analyzing shapes have been proposed in the literature, one of the newer approaches is to use an artificial neural network for detecting the desired properties [1, 2, 7].

From the cyber security standpoint this increased automation means increased attack surface. Although cyber operations against computers, networks, and data required for administering medical care are prohibited under international law, there has been a steady increase in attacks against them [15]. Awareness has an important role in cyber security of the healthcare sector, as stated by Rajamäki et al. [14]: *"The highest concern for healthcare organizations is the employee negligence followed by the fear of a cyber-attack."* In their study, Spanakis et al. analyzed cyber security in the healthcare domain and stated the fact that growth of technology utilized in healthcare concurrently increases the attack surface and thus the risk of cyber incidents increases [18]. One possible motivation can be the capability

of claiming ransoms. Modifying the automated diagnosis capability with a cyber attack may affect the treatment and in the worst case scenario lead to loss of human lives. That also raises the possibility of targeted attacks against a particular person. In conclusion, such attacks may lead to global lack of trust in automated diagnosis systems [17].

1.2 Adversarial attacks

Goodfellow et al. define adversarial examples as samples where an adversary makes a small but well-chosen perturbation into an input sample causing the artificial neural network classifier to misclassify that sample with high confidence [5]. If an adversary possesses a fast way of generating the adversarial examples, they can be used to mount various types of attacks against systems that utilize neural networks for classification tasks.

Papernot et al. state that models of machine learning are vulnerable to modified (malicious) inputs and on that account they introduced a black-box attack against deep neural networks without knowledge of the classifier training data or model [11]. Such attack methods have been introduced in some real world scenarios, for example, Stokes et al. [19] studied the attack, and furthermore, defence of malware detection models and image modifications against artificial intelligence (AI) based computer vision capabilities has been researched in [6]. When attacking against computer vision and image based machine learning, pixel modification is an evident possibility. Lin et al. tested adversarial attacks by modifying critical pixels of the image with limitations for the number of modified pixels. They changed as few as five pixels in the general CIFAR-10 dataset using a gradient-based dual iterative fusion method [8]. A short survey of model fooling attacks in the medical domain by Sipola et al. shows that at least adversarial images and patches have been used in experiments [17]. One-pixel attack is a more advanced method, in which only one pixel of an image is modified in order to fool the classifier [20]. Additionally, mitigation capabilities have been developed, Paul et al. [12] introduce mitigation of adversarial attacks on medical image systems with the conclusion that its effectiveness can be decreased by adding adversarial images in the training set. In addition to this kind of robust optimization, Xu et al. [28] mention the possibility of gradient masking and attack detection before forwarding the images to the actual classifier. However, the detection of attacks is beyond the scope of our study.

In this article we show that creating adversarial examples in the context of medical imaging is both feasible and fast. Furthermore, we show that a particular type of one-pixel perturbation is sufficient to alter legitimate input images in a fashion that causes the classifier to misclassify them with high confidence. This work is the implementation of the conceptual attack framework of using one-pixel attacks against medical imaging [16]. In addition, we explore how to create fake images that appear authentic also to the human observer. Thirdly, we present a method for altering existing images to achieve the goal. We opted to use full slide microscopy images, as they are a major target for automated analysis and digital pathology research, as well as being readily available for scientific use. Although the proposed methods are presented in a cyber attack context, the results can be used to improve or assess classifiers outside this context.

2 METHODS AND EXPERIMENTAL SETUP

Two attacks were performed: mitosis-to-normal, where the objective is to minimize the confidence score value, and normal-to-mitosis, where the objective is to maximize the confidence score value. The former attack type alters an image containing abnormal mitosis into one where the classifier fails to detect this with high confidence. The latter converts an image with normal mitosis into one where the classifier misclassifies it being an abnormal mitosis.

2.1 Attack target

The dataset used in this research is from the Tumor Proliferation Assessment Challenge 2016 (TUPAC16) [9, 25]. The dataset consists of 500 whole slide light microscopy images with known tumor proliferation scores, ground truth labels for the training set, as well as region of interest location data for 148 images. The dataset was preprocessed by a script from IBM CODAIT Center for Open-source Data & AI Technologies' *deep-histopath* repository,[1] which split the whole slide image into 64-by-64 pixel PNG-format images. The images were marked either 'mitosis' or 'normal' according to the provided labeling.

The chosen target classifier was IBM CODAIT's MAX breast cancer detector, which is here used as a pretrained black-box neural network classifier [3]. This classifier was chosen for its high ranking in the TUPAC16 challenge, and the open source nature of the code. Due to the nature of artificial neural network -based classifiers, this attack method is likely to work on other TUPAC16 contest entries as well. The obtained results do not suggest a particular failure or error in the IBM CODAIT's work or approach.

To simulate a black-box attack situation, the artificial neural network is queried through a HTTP API. Only the input image and the confidence score of the artificial neural network model for the input image are known. Inference on the model was performed by converting the image to a byte string and querying the model API residing in a Docker container. The response from the API returned a confidence score for the image. The images were also filtered based on the confidence score provided by the artificial neural network. A 'mitosis' labeled image with confidence score below 0.9 and 'normal' labeled image with score above 0.1 are filtered out of the experiment. This way the attacks focus on the unambiguous cases that should be classified correctly by the artificial neural network. Computation time was capped at five days. Consequently, 5,343 'mitosis' and 80,725 'normal' labeled images are tested using this method.

2.2 Attack outline

The goal of the attack is to find a method capable of perturbing legitimate input images in a way that causes the classifier to misclassify them with high confidence. It is usually in the interest of the attacker to find a perturbation that alters the original image as little as possible. The so-called *one-pixel attack* is achieved when the perturbation that causes a misclassification consists of altering just one pixel in the input image. To a human observer the difference between the original and altered image might be indistinguishable. As stated, two attacks are performed: mitosis-to-normal, where the objective is to minimize the confidence score value and normal-to-mitosis, where the objective is to maximize the confidence score value.

To carry out a black-box attack the adversary needs to make perturbations to the original image, and observe how the classifier under attack reacts. Su et al. proposed a method capable of creating one-pixel perturbations using differential evolution [21]. Differential evolution is an optimization method [4, 13] which can be leveraged for iteratively refining the chosen perturbations until the attacker achieves the desired misclassification confidence. In this study, we used the implementation of differential evolution in the SciPy library [27].

A color digital image can be presented as a grid of pixels, where each pixel is a mix of red, green, and blue colors, corresponding to the color sensing cells in human eye. A one-pixel perturbation can be represented by a vector: $\mathbf{x} = (x, y, r, g, b)$, where x and y are the pixel coordinates and r, g, b are the red, green and blue values of the color. All these variables are integers. The bounds for coordinates are [0,63] and the bounds for color values are [0,255].

---

[1]Available online at: https://github.com/CODAIT/deep-histopath

The initial population consists of 200 one-pixel perturbation attack vectors, the vector values are initialized using Latin hypercube sampling, which ensures that each coordinate and color value is uniformly sampled inside its bounds. A larger initial population was found to increase attack success only in some rare cases, while it slowed down attack vector search considerably due to higher computation costs. The mutation factor was set at 0.5 and the recombination factor at 0.7. Larger mutation factor and lower recombination factor values were not found to impact the attack success rate in neither mitosis-to-normal nor normal-to-mitosis attacks. Maximum iterations for the evolution were set at 100, although in practice the evolution converged on average at 44 iterations in mitosis-to-normal and 39 on normal-to-mitosis attacks. After the initial population is created, the members of the population are iterated over. The strategy for creating trial vectors was chosen as 'best1bin'.

2.3 Attack success metric

The first criterion for a successful attack is the number of steps in its evolution progress. Attacks that converged after the iteration of the initial population were found to not alter the confidence score at all or by very little margin. Thus, a successful attack needs to iterate the population more than once. On the other hand, the less iterations a success uses, the less computing power is used to find it.

The second criterion is the confidence score threshold used to determine if an attack is a success. The closer the model's confidence score is to 1, the more sure the model is that the image should be labeled 'mitosis' and the closer the score is to 0, the image is to be labeled 'normal'. To define attacks as successful, mitosis and normal attacks should reach at least 0.5 score threshold, reducing the neural network's prediction into a coin flip. If a mitosis-to-normal attack manages to lower confidence score to 0.05 or a normal-to-mitosis attack the score to 0.95, the model is fooled to predict the opposite label with high certainty.

3 RESULTS

The one-pixel attack was performed on 5,343 'mitosis' labeled images and 80,725 'normal' labeled images. The attack results were documented in a comma-separated values (CSV) file, including the name of the images used in the experiment, differential evolution parameters, original confidence score and the score after the attack. The confidence score indicates how confident the artificial neural network is that the image contains mitosis activity. The score varies between $[0, 1]$, where 0 means that the image is considered normal and 1 means that it is considered to contain mitoses.

3.1 Failures due to early convergence

Attacks where evolution converged immediately after the initial population can be considered as failed attacks. In mitosis-to-normal attacks, in 1,594 or approximately 30% of the attacks the algorithm converged already after the calculation of initial population function values, while in normal-to-mitosis attacks, 80,520 or 99.7% converged after the initial population. This is due to the tolerance value set at 0.01 and the standard deviation of the initial population being too low compared to tolerance value multiplied by mean of initial population. Lowering of the tolerance value had no impact on finding more successful populations. The tolerance value and the convergence check cause the amount of 'normal' labeled images processed to be higher than 'mitosis' labeled, because more evolution steps were performed on 'mitosis' labeled images. If the evolution converges after the initial population values, in mitosis-to-normal attacks the attack yields only 0.06 change in confidence score on mean and in normal-to-mitosis attacks the confidence score change is 0.001 on mean.
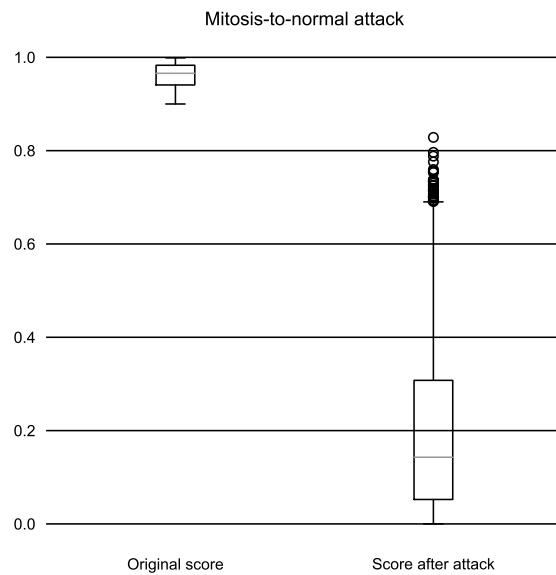
Fig. 1. Box plot visualization of mitosis-to-normal attack experiment confidence scores. The majority of the attacks were successful, lowering the confidence score below 0.5 in 3,407 or approximately 91% of the attacks. 895 or approximately 24% of the attacks manage to lower the confidence score below 0.05.

3.2 Confidence scores

The changes in the confidence score were noticeable in both attack types. On mitosis-to-normal attack, 3,407 attacks (91%) out of 3,749 managed to lower the artificial neural network's confidence score below 0.5 and 895 attacks (24%) lowered it below 0.05. On normal-to-mitosis attacks, neural network's confidence score was raised higher than 0.5 on 173 out of 205 attacks (84%) but none of the attacks managed to cross above the 0.95 score threshold.

When looking at attacks where the differential evolution algorithm did not converge on the initial population, the median confidence score difference between the original score and score after attack reaches 0.81. Applying the same filter as in mitosis-to-normal attacks, the median confidence score difference in normal-to-mitosis attacks between original images and adversarial images reaches 0.27.

Mitosis-to-normal attacks were successful in finding adversarial examples. Figure 1 has its center line at the median value, its box limits extending from 25% to 75%, its whiskers from the edges of the box to no more than 1.5 times interquartile range, ending at the farthest point in the interval and its outliers plotted as dots. The figure shows how the neural network's confidence score before the attack is on average 0.96, the maximum score is 0.99 and minimum is 0.90. After attacking the images and finding adversarial images, artificial neural network's confidence score median values are 0.1, they also reach a minimum of 0.0001 and a maximum of 0.83. The standard deviation for the scores is 0.18 and the mean is 0.20. This information is also conveyed in Table 1.

Normal-to-mitosis attacks were also successful. Before the images are attacked, neural network's confidence scores are on average 0.048, where the minimum is 0.0036 and maximum is 0.099. Figure 2 shows this as box plot, which shares its statistical characteristics with 1. After attacking the images, neural network's confidence score median is 0.31, the

Table 1. Confidence score statistics for mitosis-to-normal attack, where the number of attacks is 3,749.

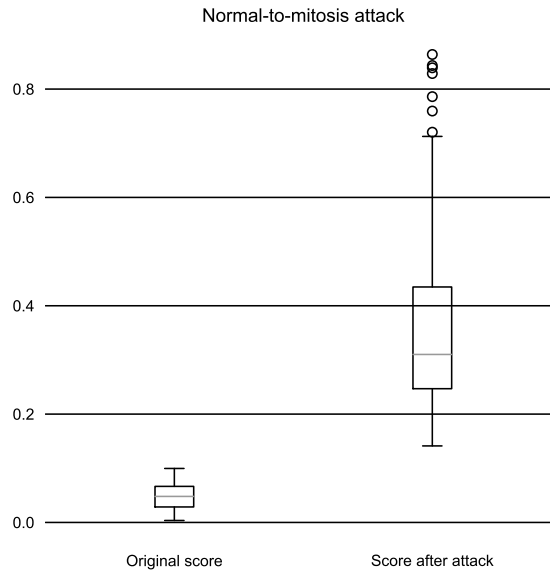|	Before attack	After attack
Maximum	0.99	0.83
Mean	0.96	0.20
Median	0.96	0.14
Standard deviation	0.02	0.18
Minimum	0.90	0.00011

Normal-to-mitosis attack

Fig. 2. Box plot visualization of normal-to-mitosis attack experiment confidence scores, where 173 or approximately 84% of attacks manage to raise the artificial neural network's confidence score above 0.5. None of the attacks manage to cross above the 0.95 score threshold.

scores minimum reaches 0.14 and maximum 0.86. The standard deviation for the scores is 0.15 and the mean is 0.36. This information is also conveyed in Table 2.

3.3 Adversarial examples

As an example, we showcase two successful one-pixel attacks. Firstly, Figure 3 shows this adversarial example that deceives the predictor to think that an image containing mitosis is a normal picture without any signs of disease. In the attacks, the most common pixel color was pure yellow, meaning RGB values (255, 255, 0), which was used in 2,214 attacks. In 122 attacks the pixel color was pure white, meaning RGB values (255, 255, 255), which was used in 122 attacks. In the rest of the attacks the pixel colors were yellow with a slightly higher blue value. Secondly, Figure 4 shows an adversarial example that deceives the predictor to think that a picture of normal cell activity contains mitosis. The most common pixel color RGB values was pure yellow (255, 255, 0) and the second most common was pure

Table 2. Confidence score statistics for normal-to-mitosis attack, where the number of attacks is 205.

	Before attack	After attack
Maximum	0.099	0.86
Mean	0.048	0.36
Median	0.048	0.31
Standard deviation	0.025	0.15
Minimum	0.0036	0.14

Label: mitosis
Confidence score 0.9074

Mitosis-to-normal attack
Confidence score 0.0305
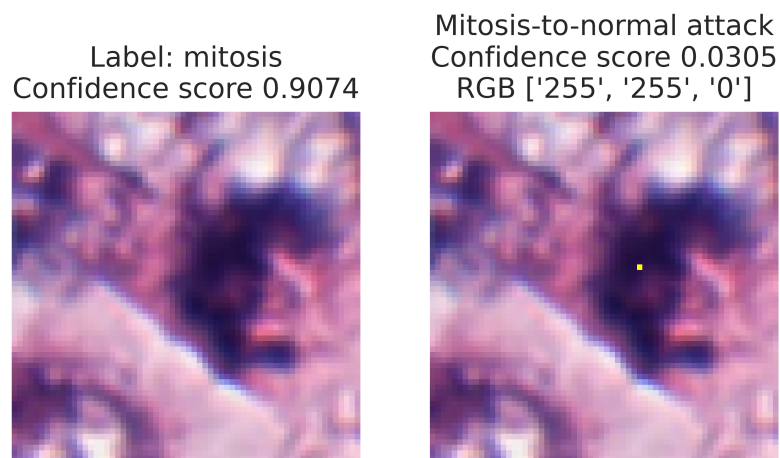RGB ['255', '255', '0']

Fig. 3. An adversarial example that is misclassified as normal even though in reality the source image is labeled as having mitosis activity. Notice the bright yellow pixel inside the dark area in the middle right part of the image.

white (255, 255, 255) and the third was pure black (0, 0, 0). There is a larger variety of colors in attack vectors than in mitosis-to-normal-attacks, but this is most likely explained due to the low amount of successful attacks.

We provide the evolutionary convergence plots for both of the example images. A convergence plot shows the change of the classifier's confidence score over the evolutionary iterations. It should be noted that the direction of the evolution depends on the type of the attack. Figure 5 shows the progress of differential evolution for the adversarial image shown in Figure 3. The lowest confidence score already reaches to almost 0.5 during the initial population attacks and drops down below 0.1 in a few steps. The minimum is reached after 40 steps. Figure 6 shows the progress of differential evolution for the image and an adversarial image in Figure 4. The maximum score of the initial population attacks is still quite low, below 0.4, but the maximum score of the population quickly rises to near 0.8 in 10 steps. The maximum 0.84 score is reached after 30 steps of the differential evolution algorithm.

4 DISCUSSION

This research demonstrates that one-pixel attacks are successful against artificial neural network diagnosis of mitosis images. It shows that a machine learning model can perform acceptably with the training and testing sets but fails catastrophically when an adversarial example is used as input. In this case, the adversarial example differs by only one

Image label: normal
Confidence score 0.0114

Normal-to-mitosis attack
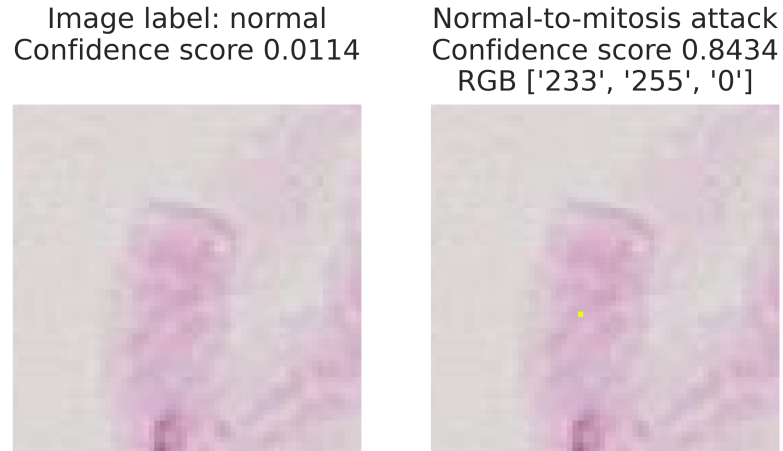Confidence score 0.8434
RGB ['233', '255', '0']

Fig. 4. An adversarial example that is misclassified as mitosis even though in reality the source image is labeled as having no mitosis activity. Notice the yellow-lime dot in the middle of the image.
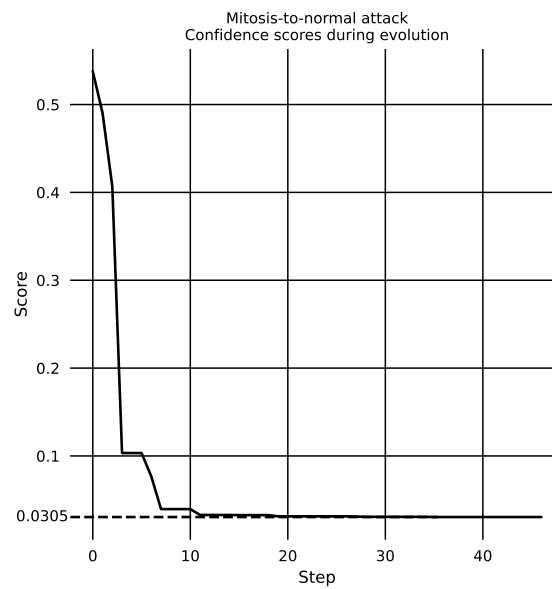

Fig. 5. Lowest neural network confidence scores during steps of differential evolution. Example of of an attack against one image, in this case the same as in Figure 3.

pixel. This hilights the need of ensuring the robustness of these artificial neural network models. While it is evident that the model works as expected in the common case, data reproduction and transmission errors, as well as cyber attacks of only one pixel, could produce undesirable results.
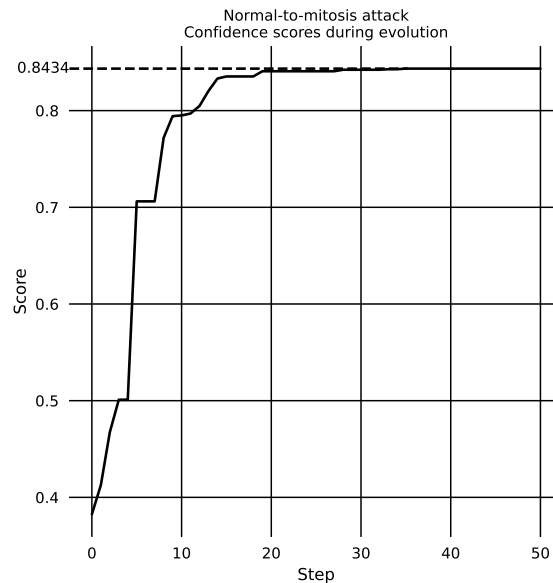
Fig. 6. Highest neural network confidence scores during steps of differential evolution. Example of of an attack against one image, in this case the same as in Figure 4.

It is evident that the attack against mitosis images is the easier one. These images might be of a more varied nature than the normal tissue images. Because of this, modifying the mitosis images does not create as considerable a change as when modifying normal tissue images. On the other hand, deceiving the artificial neural network with modified normal images was more difficult. We speculate that the neural network has likely learned to classify images with large black blobs as mitosis, thus the neural network is not easily fooled to change labels by only modifying one pixel. Larger modification of the input image would be needed for higher normal-to-mitosis attack success rate.

This result of complete reversal of classification should not be taken as a discouragement of the use of automated diagnosis systems as part of medical imaging. Instead, it shows that the medical models built using modern artificial neural network technologies can be vulnerable to unexpected attacks and other changes in the input images. Attack methods will keep evolving. At the moment, the attack pixel may be somewhat prominent, which makes their detection easy, although such artifacts could be introduced just before the analysis. On the other hand, defending against such perturbations should be easy. In the future, attack-side research ideas could include blending the attack vector color values as seamlessly to the surrounding pixels as possible, thus fooling human observers. Neverhteless, detection of such attacks and robustness of image classifiers are also important research topics.

ACKNOWLEDGMENTS

REFERENCES

[1] Md Zahangir Alom, Theus Aspiras, Tarek M. Taha, Vijayan K. Asari, T.J. Bowen, Dave Billiter, and Simon Arkell. 2019. Advanced deep convolutional neural network approaches for digital pathology image analysis: A comprehensive evaluation with different use cases. *arXiv preprint arXiv:1904.09075* (2019).

[2] Kaustav Bera, Kurt A. Schalper, David L. Rimm, Vamsidhar Velcheti, and Anant Madabhushi. 2019. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology* 16, 11 (2019), 703–715.

[3] Mike Dusenberry and Fei Hu. 2018. Deep Learning for Breast Cancer Mitosis Detection. https://github.com/CODAIT/deep-histopath/raw/master/docs/tupac16-paper/paper.pdf.

[4] Vitaliy Feoktistov. 2006. *Differential evolution.* Springer.

[5] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations.* http://arxiv.org/abs/1412.6572

[6] X. Kang, B. Song, X. Du, and M. Guizani. 2020. Adversarial Attacks for Image Segmentation on Multiple Lightweight Models. *IEEE Access* 8 (2020), 31359–31370. https://doi.org/10.1109/ACCESS.2020.2973069

[7] Pegah Khosravi, Ehsan Kazemi, Marcin Imielinski, Olivier Elemento, and Iman Hajirasouliha. 2018. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine* 27 (2018), 317–328.

[8] B. C. Lin, H. J. Hsu, and S. K. Huang. 2020. Testing Convolutional Neural Network using Adversarial Attacks on Potential Critical Pixels. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC).* 1743–1748. https://doi.org/10.1109/COMPSAC48688.2020.000-3

[9] Medical Image Analysis Group Eindhoven (IMAG/e). 2016. Tumor Proliferation Assessment Challenge 2016. http://tupac.tue-image.nl/node/3.

[10] Haidy Nasief, Cheng Zheng, Diane Schott, William Hall, Susan Tsai, and Beth Erickson amd X. Allen Li. 2019. A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer. *npj Precision Oncology* 3, 25 (2019). https://doi.org/10.1038/s41698-019-0096-z

[11] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical Black-Box Attacks against Machine Learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (Abu Dhabi, United Arab Emirates) *(ASIA CCS '17).* Association for Computing Machinery, New York, NY, USA, 506–519. https://doi.org/10.1145/3052973.3053009

[12] R. Paul, M. Schabath, R. Gillies, L. Hall, and D. Goldgof. 2020. Mitigating Adversarial Attacks on Medical Image Understanding Systems. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI).* 1517–1521. https://doi.org/10.1109/ISBI45749.2020.9098740

[13] Kenneth V Price. 2013. Differential evolution. In *Handbook of Optimization.* Springer, 187–214.

[14] J. Rajamäki, J. Nevmerzhitskaya, and C. Virág. 2018. Cybersecurity education and training in hospitals: Proactive resilience educational framework (Prosilience EF). In *2018 IEEE Global Engineering Education Conference (EDUCON).* 2042–2046. https://doi.org/10.1109/EDUCON.2018.8363488

[15] Michael N Schmitt. 2017. *Tallinn manual 2.0 on the international law applicable to cyber operations.* Cambridge University Press.

[16] Tuomo Sipola and Tero Kokkonen. 2021. One-Pixel Attacks Against Medical Imaging: A Conceptual Framework. In *Trends and Applications in Information Systems and Technologies (Advances in Intelligent Systems and Computing, Vol. 1365),* Álvaro Rocha, Hojjat Adeli, Gintautas Dzemyda, Fernando Moreira, and Ana Maria Ramalho Correia (Eds.). Springer International Publishing, Cham, 197–203. https://doi.org/10.1007/978-3-030-72657-7_19

[17] Tuomo Sipola, Samir Puuska, and Tero Kokkonen. 2020. Model Fooling Attacks Against Medical Imaging: A Short Survey. *Information & Security: An International Journal* 46, 2 (2020), 215–224. https://doi.org/10.11610/isij.4615

[18] E. G. Spanakis, S. Bonomi, S. Sfakianakis, G. Santucci, S. Lenti, M. Sorella, F. D. Tanasache, A. Palleschi, C. Ciccotelli, V. Sakkalis, and S. Magalini. 2020. Cyber-attacks and threats for healthcare – a multi-layer thread analysis. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC).* 5705–5708. https://doi.org/10.1109/EMBC44109.2020.9176698

[19] J. W. Stokes, D. Wang, M. Marinescu, M. Marino, and B. Bussone. 2018. Attack and Defense of Dynamic Analysis-Based, Adversarial Neural Malware Detection Models. In *MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM).* 1–8. https://doi.org/10.1109/MILCOM.2018.8599855

[20] J. Su, D. V. Vargas, and K. Sakurai. 2019. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation* 23, 5 (2019), 828–841. https://doi.org/10.1109/TEVC.2019.2890858

[21] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* 23, 5 (2019), 828–841.

[22] U.S. National Cancer Institute at the National Institutes of Health (NIH). [n.d.]. Cancer Incidence Rate. https://seer.cancer.gov/statistics/types/incidence.html.

[23] U.S. National Cancer Institute at the National Institutes of Health (NIH). [n.d.]. Cancer Statistics. https://www.cancer.gov/about-cancer/understanding/statistics.

[24] P J van Diest, E van der Wall, and J P A Baak. 2004. Prognostic value of proliferation in invasive breast cancer: a review. *Journal of Clinical Pathology* 57, 7 (2004), 675–681. https://doi.org/10.1136/jcp.2003.010777

[25] Mitko Veta, Yujing J. Heng, Nikolas Stathonikos, Babak Ehteshami Bejnordi, Francisco Beca, Thomas Wollmann, Karl Rohr, Manan A. Shah, Dayong Wang, Mikael Rousson, Martin Hedlund, David Tellez, Francesco Ciompi, Erwan Zerhouni, David Lanyi, Matheus Viana, Vassili Kovalev, Vitali Liauchuk, Hady Ahmady Phoulady, Talha Qaiser, Simon Graham, Nasir Rajpoot, Erik Sjöblom, Jesper Molin, Kyunghyun Paeng, Sangheum Hwang, Sunggyun Park, Zhipeng Jia, Eric I-Chao Chang, Yan Xu, Andrew H. Beck, Paul J. van Diest, and Josien P.W. Pluim. 2019. Predicting

breast tumor proliferation from whole-slide images: The TUPAC16 challenge. *Medical Image Analysis* 54 (2019), 111–121. https://doi.org/10.1016/j.media.2019.02.012

[26] Mitko Veta, Paul J. Van Diest, Mehdi Jiwa, Shaimaa Al-Janabi, and Josien P.W. Pluim. 2016. Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. *PloS one* 11, 8 (2016), e0161286.

[27] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[28] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K. Jain. 2020. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing* 17, 2 (2020), 151–178. https://doi.org/10.1007/s11633-019-1211-x

[29] Wei Zhang, Jeremy Chien, Jeongsik Yong, and Rui Kuang. 2017. Network-based machine learning and graph theory algorithms for precision oncology. *npj Precision Oncology* 1, 25 (2017). https://doi.org/10.1038/s41698-017-0029-7