One-pixel attack deceives automatic detection of breast cancer

Joni Korpihalkola[a], Tuomo Sipola[a,*], Samir Puuska[b], Tero Kokkonen[a]

[a]*Institute of Information Technology, JAMK University of Applied Sciences, PO Box 207, FI-40101 Jyväskylä, Finland*
[b]*Faculty of Information Technology, University of Jyväskylä, PO Box 35, FI-40014 Jyväskylä, Finland*

Abstract

In this article we demonstrate that a state-of-the-art machine learning model predicting whether a whole slide image contains mitosis can be fooled by changing just a single pixel in the input image. 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 diagnostic and treatments. In this research one-pixel attack is demonstrated in a real-life scenario with a real tumor dataset. The results indicate that a minor one-pixel modification of a whole slide image under analysis can affect the diagnosis. 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.

Keywords: adversarial examples, cyber security, machine learning, medical imaging, breast cancer, model safety

1. Introduction

Cancer, in its various forms, is one of the leading causes of death in the western world. A 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

---

*Corresponding author: Tel.: +358 50 310 3339;

Email addresses: joni.korpihalkola@jamk.fi (Joni Korpihalkola), tuomo.sipola@jamk.fi (Tuomo Sipola), sapepuus@student.jyu.fi (Samir Puuska), tero.kokkonen@jamk.fi (Tero Kokkonen)

cancers per 100,000 population (U.S. National Cancer Institute at the National Institutes of Health (NIH), a). 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 (U.S. National Cancer Institute at the National Institutes of Health (NIH), b). The institute stated that *"Cancer is among the leading causes of death worldwide. In 2018, there were 18.1 million new cases. The number of new cancer cases per year is expected to rise to 29.5 million"* (U.S. National Cancer Institute at the National Institutes of Health (NIH), b).

The high number of new incidences (high CIR) means that cancer and various cancer-related medical tasks require substantial time and resources. Although cancer is not one disease, many laboratory diagnostic methods are shared between various types. These include morphologic methods where microscopy in combination with various staining methods is used to draw conclusions based on various properties and counts of cells in a biological sample. In many common types of cancers, early detection is a key factor in improving the prognosis (van Diest et al., 2004). Since detection plays a major role in patient outcomes, automating some of this work using techniques such as machine learning can lead to faster detection, increased throughput, and reduced costs.

From the cyber security standpoint this increased automation means increased attack surface. Disrupting the operation of nation's critical infrastructure has been an integral part of nation-state level attacks. The goal of these attacks may simply be to erode trust in the nation's capability to provide services to citizens, or in the worst case, be a part of an armed conflict. Healthcare sector is a major part of critical infrastructure, and as such a target for advanced cyber operations. 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 (Schmitt, 2017). Awareness has an important role in cyber security of the healthcare sector, as stated by Rajamäki et al. (2018): *"The highest concern for healthcare organizations is the employee negligence followed by the fear of a cyber-attack."*

Automated or partly automatized analysis systems are priority targets for cyber attacks, as a disruption in these systems causes major effects resulting in decreased capability to diagnose and treat patients, increased expenses, and overall reduction in

trust towards automated systems. 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 (Spanakis et al., 2020).

In cancer diagnosis, computer vision and machine learning can be used to automate various tasks (Zhang et al., 2017; Nasief et al., 2019) where e.g. a count of cells needs to be made from an image (Veta et al., 2016). 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. These tools aim to make cancer diagnosis less expensive and less time consuming (Khosravi et al., 2018; Bera et al., 2019; Alom et al., 2019).

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 (Papernot et al., 2017). Such attack methods have been introduced in some real world scenarios, for example, Stokes et al. (2018) 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 (Kang et al., 2020). 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 (Lin et al., 2020). One-pixel attack is a more advanced method, in which only one pixel of an image is modified in order to fool the classifier (Su et al., 2019). Additionally, mitigation capabilities have been developed, Paul et al. (2020) 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. mention the possibility of gradient masking and attack detection before forwarding the images to the actual classifier. (Xu et al., 2020)

The healthcare sector can be seen as a valuable target for cyber attackers with different motivations. One possible motivation can be the capability of claiming ransoms. Modifying the automated diagnosis capability with a cyber attack may affect the treat-

ment 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 (Sipola et al., 2020).

In this study we attempt to construct images that fool real world state-of-the-art classifiers. 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

Morphological methods leverage various features of cells in deciding whether they have neoplastic characteristics. The cells are extracted from the tissue under study, and visually inspected using microscopy and stains. The so called whole slide images are high resolution pictures that are often digitized and viewed using computers. This format also provides a suitable basis for a more automated approach.

In this article we examine classifying methods based on artificial neural networks. An artificial neural network classifiers are usually trained using samples, i.e. images, that have a known label in a fashion that allows the trained classifier to recognize also samples that it has not previously "seen".

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 (Goodfellow et al., 2015). 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. In this article we show that creating adversarial examples in a context of medical imaging is both feasible and fast. Furthermore, we show that a particular type of perturbation is sufficient to alter the vast majority of le-

gitimate input images in a fashion that causes the classifier to misclassify them with high confidence.

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) (Medical Image Analysis Group Eindhoven (IMAG/e), 2016; Veta et al., 2019). 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 classifier was IBM CODAIT's MAX breast cancer detector (Dusenberry and Hu, 2018). 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

---

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

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 (Su et al., 2019). Differential evolution is an optimization method (Feoktistov, 2006; Price, 2013) 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 Scikit-learn library (Pedregosa et al., 2011).

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].

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'. In the strategy, two random vectors are chosen to mutate the best performing vector in the population, meaning the attack vector that achieved the lowest value from the artificial neural network output. The parameters from the best vector are mutated using the mutation factor and the difference of the two random vectors.

A trial vector is created. Random values for each parameter are generated using binomial distribution; if the random value for the parameter is lower than the recombination factor value, the mutated value in the parameter is inserted into the trial vector. If the random value is higher than the recombination factor value, the value from the best performing vector is carried over into the trial vector. However, one random value is always replaced with a mutated value, even if the binomial distribution values were all below the recombination factor value.

After the trial vector is created, its performance is tested by modifying the target image with the vector's values. If the trial vector performs better than the original member of the population, it is replaced with the trial vector. If the trial member also performs better than the best member of the population, the best member is also replaced by the trial vector.

When the attack vector population is iterated over, a convergence check is performed. The convergence check compares if the standard deviation of the population's

confidence scores is lower than the absolute mean of the population's confidence scores multiplied by tolerance factor and the absolute tolerance value is added to the result multiplication result. If the comparison is true, the algorithm converges and the best population member is the best attack vector found and its confidence score is the function's minimum value.

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.

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.

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 attacked 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
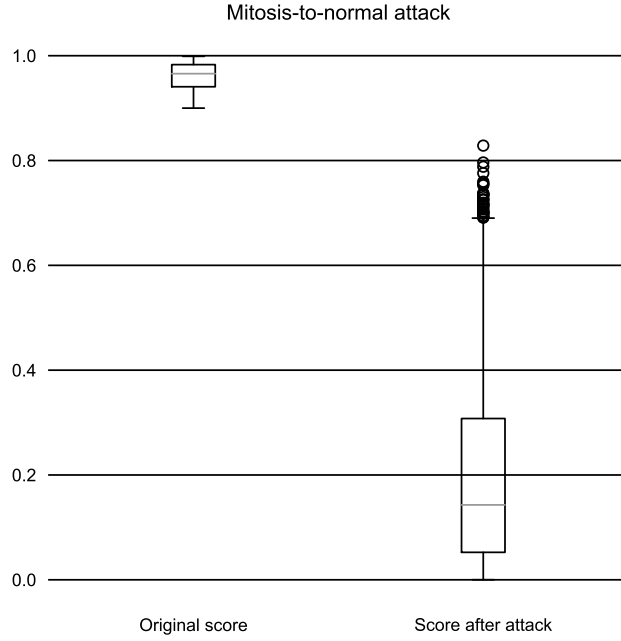
9

Figure 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.

	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	0.99	0.00011

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

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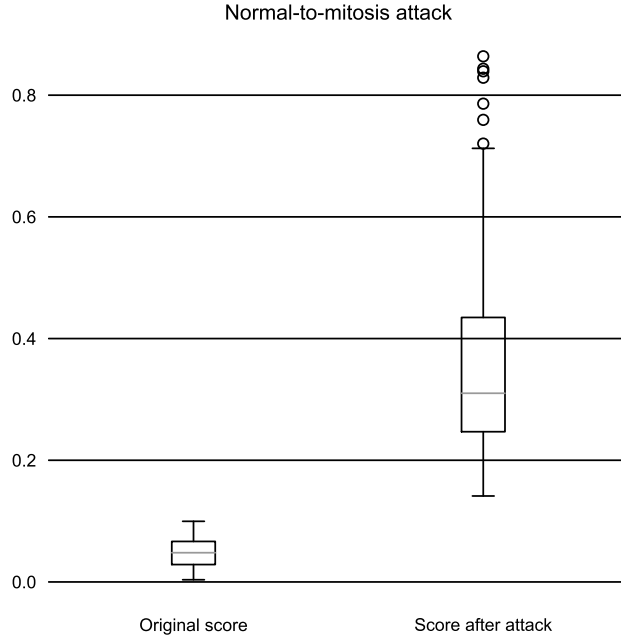
Figure 2: Box plot visualization of normal-to-mitosis attack experiment confidence scores, where 173 or approximately 84% of attacks manage to raise the artificial neural network's confidence score above 0.5. None of the attacks manage to cross above the 0.95 score threshold.

	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

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

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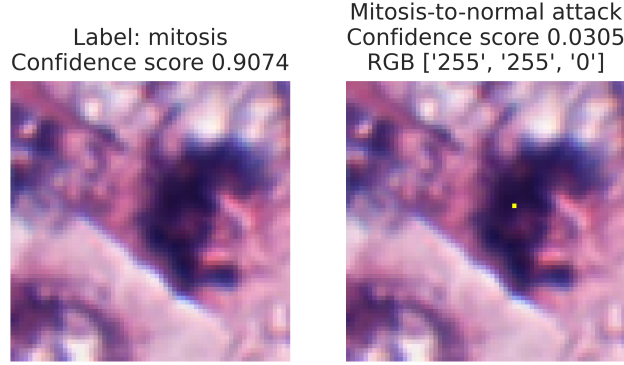
Figure 3: An adversarial example that is misclassified as normal even though in reality the source image is labeled as having mitosis activity. Notice the bright yellow pixel inside the dark area in the middle right part of the image.

3.3. Adversarial examples

As an example, we showcase two successful 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 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 images. Figure 5 shows the progress of differential evolution for the attacked 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

Image label: normal
Confidence score 0.0114

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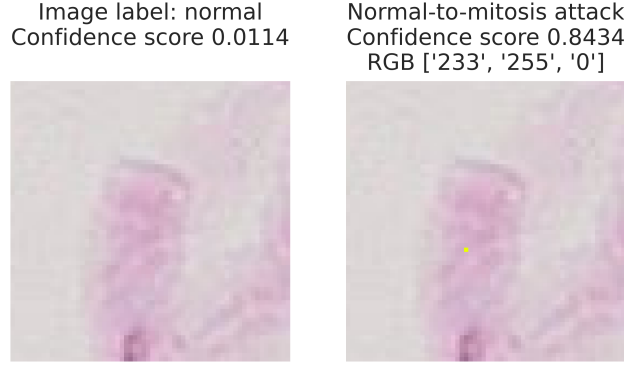Confidence score 0.8434
RGB ['233', '255', '0']

Figure 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.

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 analysis 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. 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.

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
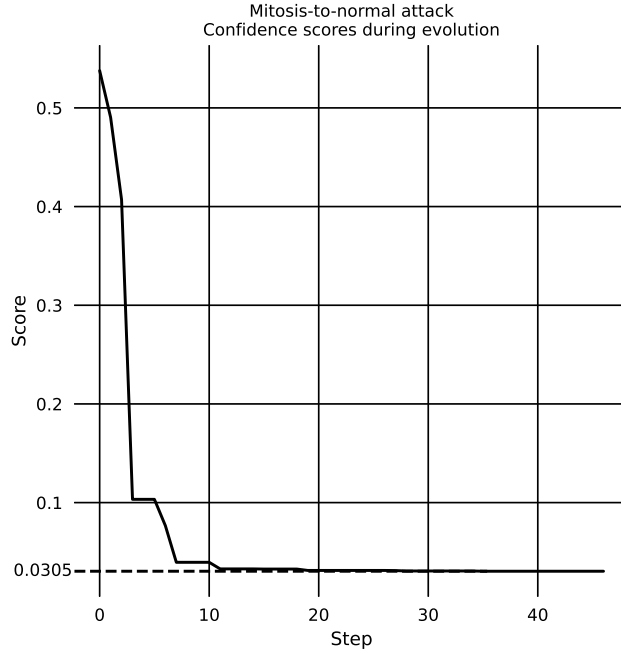
13

Figure 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.

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 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. As with all technology, its limitations should be known in order to correctly utilize the capabilities it provides.

From the cyber security point of view this should be considered as an alarming finding because motivated and skilled attackers can execute such attacks quite easily, if they have access to the medical image repository. This real-life scenario can be considered possible because there are more and more attacks against healthcare systems
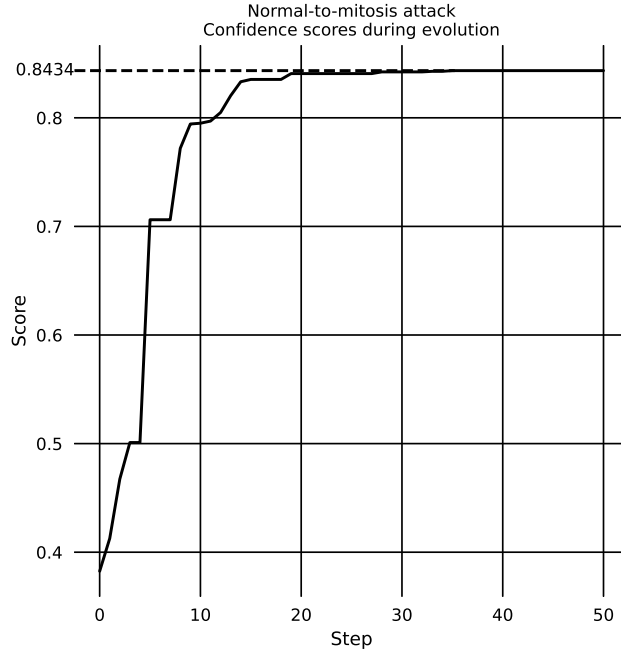
14

Figure 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.

and with such an attack scenario the motivation can be ransoms. Attackers might think that there is a high probability that ransoms will be paid when a threat of misdiagnosis of cancer exists.

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. 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.

Data availability

The whole slide images and labeling for the training dataset are available from the Tumor Proliferation Assessment Challenge 2016 (TUPAC16) website (`http://tupac.tue-image.nl/`), which derives the data from the Cancer Genome Atlas by the TCGA

Research Network (`https://www.cancer.gov/tcga`), the AMIDA13 dataset and from two different pathology centers in The Netherlands (Medical Image Analysis Group Eindhoven (IMAG/e), 2016; Veta et al., 2019; U.S. National Cancer Institute at the National Institutes of Health (NIH), c; Veta et al., 2015). Restrictions apply to the TUPAC16 dataset, as the labels are available for registered users only. However, the data are available upon reasonable request, through collaborative investigations and with permission of Medical Image Analysis Group Eindhoven (IMAG/e). Furthermore, the experiment results are available in CSV format in the same repository as the programming codes used in the experiment, available from the corresponding author upon reasonable request.

Competing interests statement

The authors declare no competing interests.

Author contributions

Joni Korpihalkola: Software, Formal analysis, Writing - Original draft, Visualization. **Tuomo Sipola:** Conceptualization, Methodology, Validation, Data curation, Writing - Original draft, Supervision. **Samir Puuska:** Conceptualization, Methodology, Data curation, Software, Writing - Original draft. **Tero Kokkonen:** Conceptualization, Writing - Original draft, Funding acquisition.

Acknowledgments

References

Alom, M.Z., Aspiras, T., Taha, T.M., Asari, V.K., Bowen, T., Billiter, D., Arkell, S., 2019. Advanced deep convolutional neural network approaches for digital pathology image analysis: A comprehensive evaluation with different use cases. arXiv preprint arXiv:1904.09075 .

Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., Madabhushi, A., 2019. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nature reviews Clinical oncology 16, 703–715.

van Diest, P.J., van der Wall, E., Baak, J.P.A., 2004. Prognostic value of proliferation in invasive breast cancer: a review. Journal of Clinical Pathology 57, 675–681. URL: `https://jcp.bmj.com/content/57/7/675`, doi:`10.1136/jcp.2003.010777`.

Dusenberry, M., Hu, F., 2018. Deep learning for breast cancer mitosis detection. https://github.com/CODAIT/deep-histopath/raw/master/docs/tupac16-paper/paper.pdf.

Feoktistov, V., 2006. Differential evolution. Springer.

Goodfellow, I., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples, in: International Conference on Learning Representations. URL: `http://arxiv.org/abs/1412.6572`.

Kang, X., Song, B., Du, X., Guizani, M., 2020. Adversarial attacks for image segmentation on multiple lightweight models. IEEE Access 8, 31359–31370. doi:`10.1109/ACCESS.2020.2973069`.

Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., Hajirasouliha, I., 2018. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. EBioMedicine 27, 317–328.

Lin, B.C., Hsu, H.J., Huang, S.K., 2020. Testing convolutional neural network using adversarial attacks on potential critical pixels, in: 2020 IEEE 44th Annual

Computers, Software, and Applications Conference (COMPSAC), pp. 1743–1748. doi:`10.1109/COMPSAC48688.2020.000-3`.

Medical Image Analysis Group Eindhoven (IMAG/e), 2016. Tumor proliferation assessment challenge 2016. `http://tupac.tue-image.nl/node/3`.

Nasief, H., Zheng, C., Schott, D., Hall, W., Tsai, S., amd X. Allen Li, B.E., 2019. A machine learning based delta-radiomics process for earlyp rediction of treatment response of pancreatic cancer. npj Precision Oncology 3. doi:`10.1038/s41698-019-0096-z`.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A., 2017. Practical black-box attacks against machine learning, in: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, Association for Computing Machinery, New York, NY, USA. p. 506–519. doi:`10.1145/3052973.3053009`.

Paul, R., Schabath, M., Gillies, R., Hall, L., Goldgof, D., 2020. Mitigating adversarial attacks on medical image understanding systems, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1517–1521. doi:`10.1109/ISBI45749.2020.9098740`.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Price, K.V., 2013. Differential evolution, in: Handbook of Optimization. Springer, pp. 187–214.

Rajamäki, J., Nevmerzhitskaya, J., Virág, C., 2018. Cybersecurity education and training in hospitals: Proactive resilience educational framework (prosilience ef), in: 2018 IEEE Global Engineering Education Conference (EDUCON), pp. 2042–2046. doi:`10.1109/EDUCON.2018.8363488`.

Schmitt, M.N., 2017. Tallinn manual 2.0 on the international law applicable to cyber operations. Cambridge University Press.

Sipola, T., Puuska, S., Kokkonen, T., 2020. Model fooling attacks against medical imaging: A short survey. Information & Security: An International Journal 46, 215–224. doi:10.11610/isij.4615.

Spanakis, E.G., Bonomi, S., Sfakianakis, S., Santucci, G., Lenti, S., Sorella, M., Tanasache, F.D., Palleschi, A., Ciccotelli, C., Sakkalis, V., Magalini, S., 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), pp. 5705–5708. doi:10.1109/EMBC44109.2020.9176698.

Stokes, J.W., Wang, D., Marinescu, M., Marino, M., Bussone, B., 2018. Attack and defense of dynamic analysis-based, adversarial neural malware detection models, in: MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM), pp. 1–8. doi:10.1109/MILCOM.2018.8599855.

Su, J., Vargas, D.V., Sakurai, K., 2019. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation 23, 828–841. doi:10.1109/TEVC.2019.2890858.

Su, J., Vargas, D.V., Sakurai, K., 2019. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation 23, 828–841.

U.S. National Cancer Institute at the National Institutes of Health (NIH), a. Cancer Incidence Rate. https://seer.cancer.gov/statistics/types/incidence.html.

U.S. National Cancer Institute at the National Institutes of Health (NIH), b. Cancer Statistics. https://www.cancer.gov/about-cancer/understanding/statistics.

U.S. National Cancer Institute at the National Institutes of Health (NIH), c. The Cancer Genome Atlas Program. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., Hedlund, M., Tellez, D., Ciompi, F., Zerhouni, E., Lanyi, D., Viana, M., Kovalev, V., Liauchuk, V., Phoulady, H.A., Qaiser, T., Graham, S., Rajpoot, N., Sjöblom, E., Molin, J., Paeng, K., Hwang, S., Park, S., Jia, Z., Chang, E.I.C., Xu, Y., Beck, A.H., van Diest, P.J., Pluim, J.P., 2019. Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. Medical Image Analysis 54, 111–121. doi:10.1016/j.media.2019.02.012.

Veta, M., Van Diest, P.J., Jiwa, M., Al-Janabi, S., Pluim, J.P., 2016. Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. PloS one 11, e0161286.

Veta, M., Van Diest, P.J., Willems, S.M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A.B., Vestergaard, J.S., Dahl, A.B., et al., 2015. Assessment of algorithms for mitosis detection in breast cancer histopathology images. Medical Image Analysis 20, 237–248. doi:10.1016/j.media.2014.11.010.

Xu, H., Ma, Y., Liu, H.C., Deb, D., Liu, H., Tang, J.L., Jain, A.K., 2020. Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing 17, 151–178. doi:10.1007/s11633-019-1211-x.

Zhang, W., Chien, J., Yong, J., Kuang, R., 2017. Network-based machine learning and graph theory algorithms for precision oncology. npj Precision Oncology 1. doi:10.1038/s41698-017-0029-7.