Chromatic and spatial analysis of one-pixel attacks against an image classifier

Janne Alatalo, Joni Korpihalkola, Tuomo Sipola, and Tero Kokkonen

Institute of Information Technology, JAMK University of Applied Sciences, Jyväskylä, Finland
{janne.alatalo,joni.korpihalkola,tuomo.sipola,tero.kokkonen}@jamk.fi

Abstract. One-pixel attack is a curious way of deceiving neural network classifiers by changing only one pixel in the input image. The full potential and boundaries of this attack method are not yet fully understood. In this research, the successful and unsuccessful attacks are studied in more detail to illustrate the working mechanisms of a one-pixel attack created using differential evolution. The data comes from our earlier studies where we applied the attack against medical imaging. We used a real breast cancer tissue dataset and a real classifier as the attack target. This research presents ways to analyze chromatic and spatial distributions of one-pixel attacks. In addition, we present one-pixel attack confidence maps to illustrate the behavior of the target classifier. We show that the more effective attacks change the color of the pixel more, and that the successful attacks are situated at the center of the images. This kind of analysis is not only useful for understanding the behavior of the attack but also the qualities of the classifying neural network.

Keywords: one-pixel attack · classification · perturbation methods · visualization · cybersecurity

1 Introduction

The use of Artificial Intelligence (AI), including sub-branches Machine learning (ML) and Deep Learning (DL), is continuously increasing as support for decision making in automated image analysis of medical imaging [20,6]. One enabler for such evolution is that there is the abundance of available data for research and development activities in the medical domain [11]. However, from the cyber security standpoint, this evolution fosters attack surface, and it should be realized that new technologies attract malicious actors and especially medical domain can be seen as a valuable target to gain profit by causing disruptions. It is noticeable that most of the medical data has sensitive nature. For example, Europol has announced that during the ongoing COVID-19 pandemic, the pandemic-themed cybercrime activities and campaigns are also targeted to healthcare organizations. [15]. Newaz et al. propose an adversarial attack against ML enabled smart healthcare system [9]. Attacks against new technologies might induce harmful

The original article appeared as: Janne Alatalo, Joni Korpihalkola, Tuomo Sipola and Tero Kokkonen. "Chromatic and spatial analysis of one-pixel attacks against an image classifier." In: Networked Systems. NETYS 2022. Ed. by Mohammed-Amine Koulali and Mira Mezini. Vol. 13464. Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2022, pp. 303–316. DOI: 10.1007/978-3-031-17436-0_20

effects: considerable time to recover, mistrust against AI-based models and even fear of misdiagnosis. It is noticeable that Internet of Things (IoT) devices have a remarkable role in the healthcare [2] and there are known security issues with IoT. Several AI models are in risk for adversarial attacks [19] Liu et al. introduce and summarize the DL associated attack and defense methods [7], while Qayyum et al. [10] introduce methods to warrant secure ML for healthcare. Integrity and unauthorized usage of medical image data is important when considering attacks against AI based medical imaging. In that sense, Kamal et al. proposed image encryption algorithm for securing medical image data [3].

One-pixel attack is an adversarial method that changes just one pixel in an image to cause misclassification. The attack is created by using optimization to find the best pixel that flips the classification decisiton made by a classifier [14]. However, its sensitivity to change and effectiveness are not fully understood. A few methods have been proposed to visualize the effect of one-pixel attacks. Wang et al. propose pixel adversarial maps and probability adversarial maps [18]. Vargas et al. go further, and use internal information from the neural network model to create propagation maps to show the influence of one-pixel attacks through convolution layers. [16]

In this study, we provide tools to understand the behavior of a neural network classifier targeted by the one-pixel attack. Our present analysis is a natural extension to our prior studies related to the attack method. Earlier, we have introduced a list of methods to fool artificial neural networks used in medical imaging [13]. One-pixel attack appeared to be a comprehensive and realistic attack vector, so we decided to further investigate it as a conceptual framework in the medical imaging domain [12]. When the concept and usability of the attack were understood, we succeeded to implement the technical one-pixel attack against real neural network models used in medical imaging [5]. That first technical attack was a success, but the pixel changes in the images were quite easily observable by a human. It seemed that the attack was not realistic or comprehensive for real-world attackers, so we decided to further develop the attack methodology [4].

The new tools we propose somewhat differ from the earlier studies. All the methods complement each other when investigating the classification effects of one-pixel changes to images. While the other methods are useful when trying to understand the internal state of the classifier (such as Vargas et al. [16]) or mapping attacks against each pixel (such as Wang et al. [18]), our confidence map approach directly addresses the classification result. Wang et al. use *successful attacks generated for each pixel* as a base for their maps [18]. Our periodicity analysis here concerns *successful attack locations of each image* that have been generated earlier. Furthermore, our confidence map analysis iterates over the color space to saturate each pixel in a brute force manner, as we do not use optimization to find attack pixels during the analysis.

The rest of the paper is organized as follows. First, data source and analysis methods, including confidence map computation, are introduced in section 2. Results of chromatic, spatial and periodicity analysis are presented in section 3

with tables and figures. Finally, the study is concluded with final discussion and future research topics in section 4.

2 Methods

2.1 Data source

In our previous publications we introduced how an artificial neural network image classifier model could be fooled by changing only one pixel in the input image [5,4]. Those studies targeted IBM CODAIT MAX breast cancer detector which uses a modified ResNet-50 model [1]. The model is an open-source convolutional neural network classifier predicting the probability that the input image contains mitosis. The previous studies used a pretrained version of the model that was trained using the TUPAC16 breast cancer dataset [8,17]. We use the same model in this research.

The study used the one-pixel attack to find adversarial images that would make the model predict wrong results for the input images [14]. This method uses differential evolution optimization, where a population of breast cancer images is attacked by randomly choosing one pixel and randomly changing the pixel's colors to new values. The color values are mutated until the lowest confidence score is achieved for the breast cancer image. The method efficiently finds possible one-pixel changes to the image that changes the prediction outcome.

The targeted model can be fooled in two ways. If the model predicts strong probability of mitosis for the input image, then the one-pixel attack is used to find the pixel that lowers the predicted mitosis probability when the pixel color is changed (*mitosis-to-normal*). The other way to fool the model is to try to increase the predicted mitosis probability when the model predicts low mitosis confidence score for some input image (*normal-to-mitosis*). The study explored both possible cases of fooling the model. The study concluded that both *mitosis-to-normal* and *normal-to-mitosis* attacks are possible, but of those two, *mitosis-to-normal* attacks are considerably easier to carry out.

The dataset used in this study contains the one-pixel attack results from the previous study [5], and information of the attacked image, such as the attack pixel's location in the image and the nearby neighboring pixels' color values of the attacked pixel. We were interested in studying attacks that were at least partially successful. We considered *normal-to-mitosis* attacks that raise the confidence score above 0.1 and *mitosis-to-normal* attacks that lowered the confidence score below 0.9 to be potentially dangerous attacks, and included all of them to our visualizations. Using these filters, 3,871 *mitosis-to-normal* attacks and 319 *normal-to-mitosis* attacks were used as a visualization dataset. Although not all attacks in this dataset were successful in flipping the classification result to other class, we consider them to be successful because they change the confidence score perceptually enough that the result is no longer trustworthy.

2.2 One-pixel attack confidence map computation

In addition to analyzing the results from our previous paper, we also carried out additional tests for some of the dataset images by brute forcing a subset of all possible attack vectors for the images, producing a one-pixel attack confidence map. This gave us a clearer view how the successful attack vectors were positioned in individual images. The brute force computation was conducted on a few handpicked images that were chosen based on our previous paper results in a way that we had successful and failed examples of both *mitosis-to-normal* and *normal-to-mitosis* attack types.

This research used color images, hence each pixel has three color channels and the color value for each channel has a value between $0-255$. This means that the total number of possible colors for a single pixel is $16,777,216$. The images were 64×64 pixels in size, so the total number of all possible attack vectors is $68,719,476,736$ for a single image. We concluded that computing all possible vectors for the images is not worthwhile; therefore, we settled on a subset of all possible attack colors. The selected set of colors C (1) was generated by taking every fifth color value for each channel and taking all their color combinations. In the equation, r, g and b are the red, green, and blue color channels:

$$C = \{(r, g, b) \mid r, g, b \in \{0, 5, 15 \dots 255\}\}. \tag{1}$$

Even when the brute forced colors were reduced to the set C, there was still $140,608$ different colors for a single pixel, meaning that the total number of attack vectors for a single image was still $575,930,368$. With that many images we could not use the Docker containerized version of the model that was used in our previous study over the HTTP API, because the containerized version of the model does not support GPU computation or image batching. We overcame this problem by deploying the model to our computation server without the containerization layer and implementing a highly efficient GPU accelerated data-pipeline that implemented the one-pixel modifications on GPU without needing to continuously copy the images between CPU and GPU memory. With this setup computing the $575,930,368$ attack vectors for one image took about 5 hours on our computation server using one Nvidia Tesla V100 GPU.

The results of the brute force attack vector analysis were reduced to minimum, maximum and average score values for each pixel coordinate.

Let $I_{x,y}$ be the set of all modified images where pixel coordinate (x, y) value is replaced with color value $c \in C$ in the image under brute force computation. Let f be the model that predicts the score for the images. The results of the brute force attacks were processed with method described in Equation 2 and Algorithm 1.

$$\begin{aligned} s_{max}(x, y) &= \max(\{f(i) \mid i \in I_{x,y}\}) \\ s_{min}(x, y) &= \min(\{f(i) \mid i \in I_{x,y}\}) \\ s_{avg}(x, y) &= \text{avg}(\{f(i) \mid i \in I_{x,y}\}) \end{aligned} \tag{2}$$

Algorithm 1 Brute force results processing algorithm

```
maxscores ← ARRAY[64][64]
minscores ← ARRAY[64][64]
avgscores ← ARRAY[64][64]
for x ← 0 to 63 do
  for y ← 0 to 63 do
    maxscores[x][y] ← s_max(x, y)
    minscores[x][y] ← s_min(x, y)
    avgscores[x][y] ← s_avg(x, y)
  end for
end for
```

3 Results

3.1 Chromatic and spatial analysis

The difference between color values of two different pixels was measured by root mean square error (RMSE).

$$h(\mathbf{x}) = \sqrt{\frac{(c_r - c_{r\mu})^2 + (c_g - c_{g\mu})^2 + (c_b - c_{b\mu})^2}{3}},$$

where c_r, c_b, c_g are the color values of the attack vector and $c_{r\mu}$, $c_{g\mu}$, $c_{b\mu}$ are the means of the attack vector's surrounding pixels' color values. All values were scaled within the range $[0, 1]$.

When the attacks managed to fool the neural network, the error function values were high in *mitosis-to-normal* attacks, as can be observed from Figure 1, which shows one vertical and one horizontal cluster. This indicates that the attacks which managed to lower confidence score the most had pixel color values noticeably different from the surrounding colors. The positioning of the attack pixel also matters, since some attacks had a higher color difference between neighboring pixels and still did not manage to lower the confidence score by more than 0.2.

In *normal-to-mitosis* attacks the error values were lower than in *mitosis-to-normal* attack, as can be seen in Figure 2, which shows no clusters; instead, the dots are more evenly distributed between the lower X axis values.

Mean, median and standard deviation numerical measures were calculated for the attacks. In the Table 1, the X and Y mean and median indicate that the attacks were mostly located at the center of the 64 by 64 pixels images. Meanwhile, the color values of red and green were near the maximum value of 255, while blue values were lower with higher standard deviation compared to red and green.

In normal images, the statistical measures listed in Table 2 show that the attack vector is mostly again located at the center of the image, while there is much greater variation in red, green and blue color values, with a standard deviation between 90 and 100 in all of them.

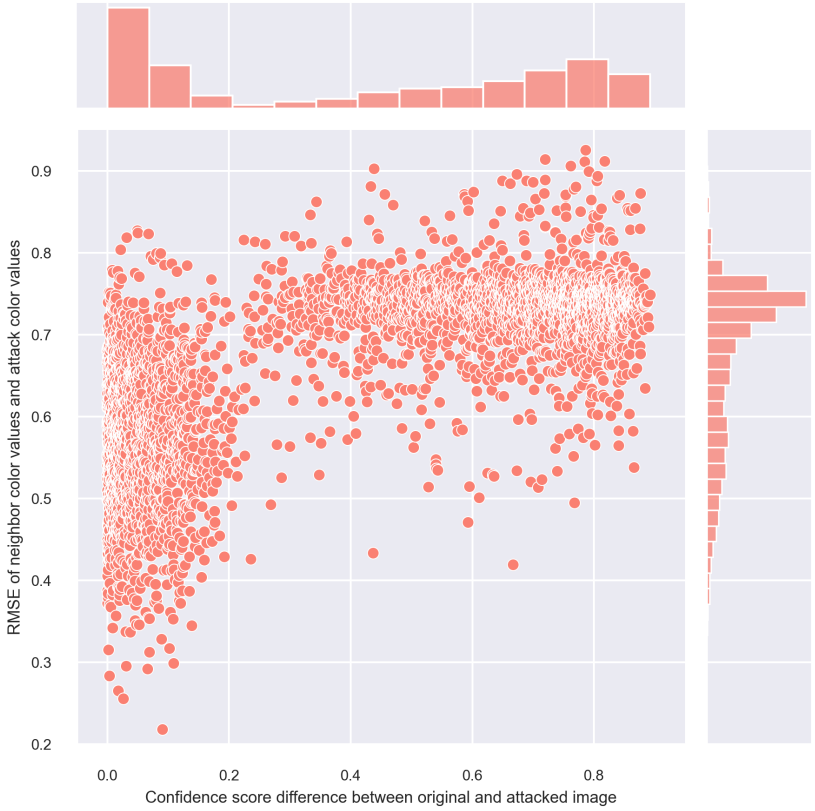


Fig. 1: Scatter plot of error function values between the attack pixel color values and neighboring pixel color values. Notice the vertical cluster at low error values and horizontal cluster at higher error values.

Table 1: Statistical measures for *mitosis-to-normal* attacks ($N = 3871$)

	X	Y	Red	Green	Blue
Mean	32.40	29.30	231.14	227.24	67.07
Median	32	30	255	255	37
SD	8.2	8.59	41.62	45.99	77.85

Table 2: Statistical measures for *normal-to-mitosis* attacks ($N = 319$)

	X	Y	Red	Green	Blue
Mean	31.55	31.15	145.28	153.31	124.29
Median	32	32	154	168	129
SD	10.54	10.77	92.62	93.04	99.53

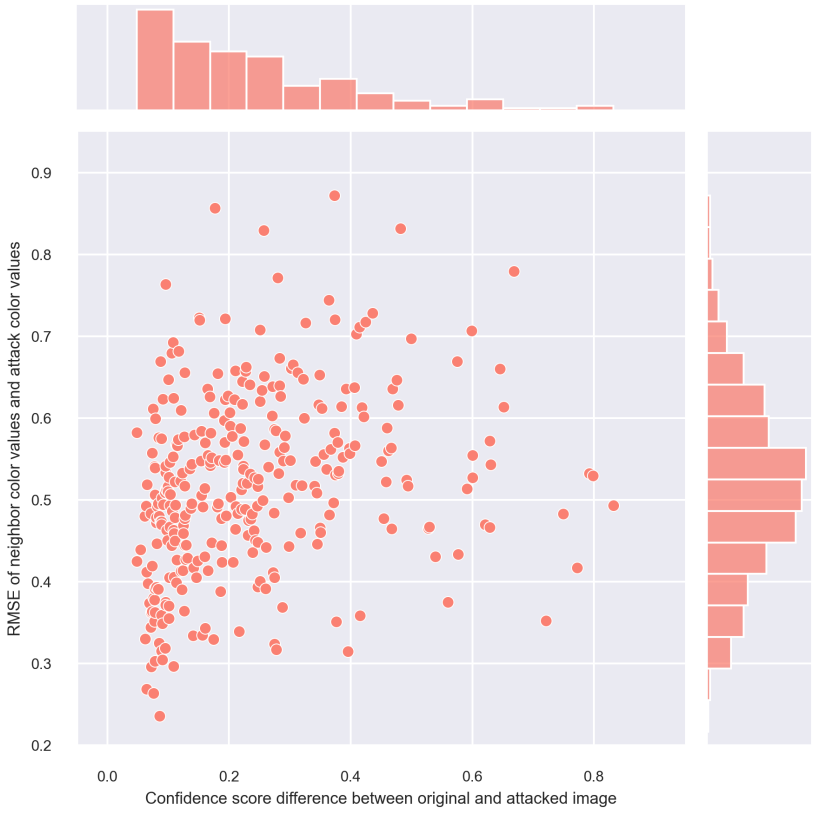


Fig. 2: Scatter plot of error function values between the attack pixel color values and neighboring pixel color values.

The statistical measures show that the dataset is most likely preprocessed in such a manner that the features used by the neural network to classify an image to either mitosis or normal class are located in the center of the image. Higher red and green color values were the key in fooling the neural network in both attacks, while blue color values were closer to zero or in the middle of the color range. In the TUPAC16 dataset, the mitosis activity was low in color range, so the neural network might be fooled by values in the higher color range.

3.2 Periodicity analysis

The targeted model is a neural network with convolution layers, which shift through the input image in smaller windows and step to the right in steps. To check for biases in the convolutional model, the best attack locations for all the target images is visualized in a heatmap in Figure 3.

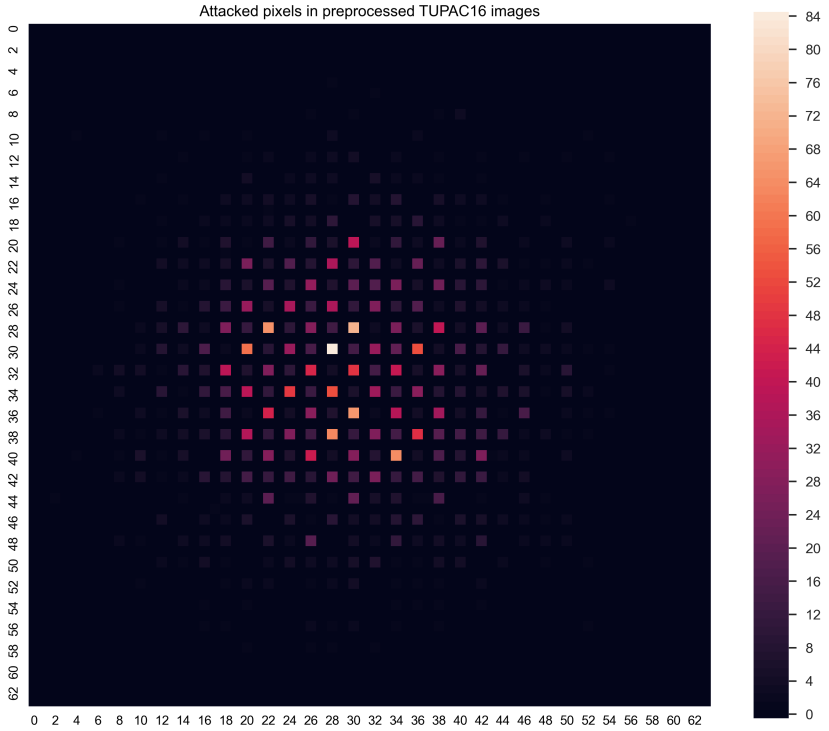


Fig. 3: A heatmap of attack placements in images. Notice the checkerboard pattern at the center.

There was a smaller ratio of successful attacks in *normal-to-mitosis* direction, and the heatmap visualization in Figure 4 does not show any significant clusters

or patterns. There is less periodicity and the center of the image is a more prominent location for the successful attacks.

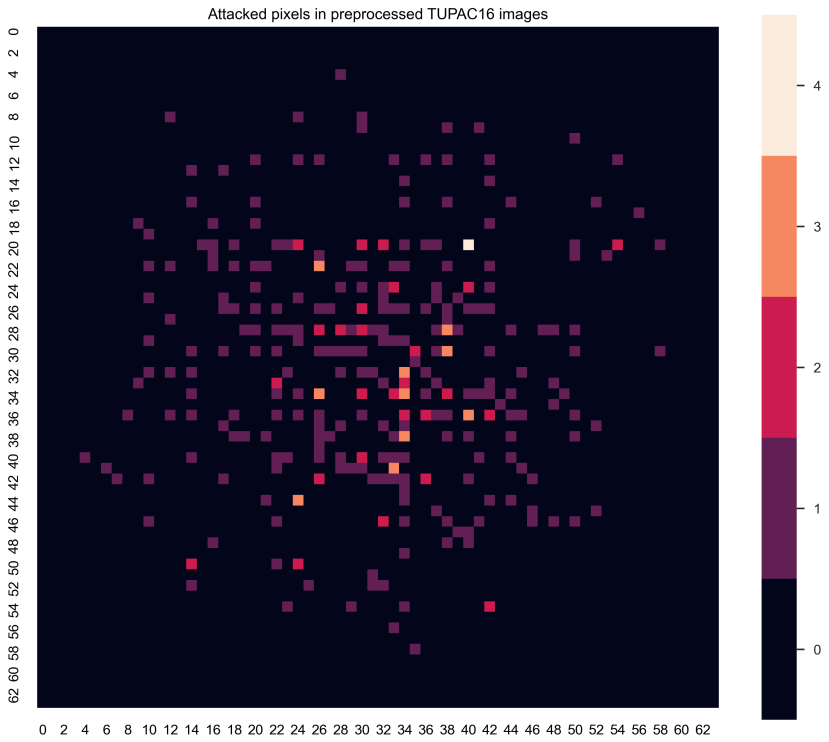


Fig. 4: A heatmap of attack locations in images. The attacks are placed mostly around the center of the image.

One of the most remarkable features of the spatial diagrams is the periodicity of the *mitosis-to-normal* attacks. Almost all successful attack pixels have coordinates with even numbers. From all of the $5,343$ *mitosis-to-normal* attacks, the differential evolution algorithm settled on pixel coordinates that had even numbers for both coordinates $5,334$ times. Only 9 times did the algorithm have best success with coordinates where both or one of the coordinates was an odd number. Only 1 of the 9 odd coordinate attack vectors was successful of lowering the score below 0.5 with modified score of 0.387.

For *normal-to-mitosis* attacks the coordinates also preferred even coordinates; however, not so clearly. From all of the $80,725$ attacks $49,573$ or 61.4% settled on even coordinates and $31,152$ or 38.6% settled on odd coordinates.

Our first reaction was to review the attack code for periodic error but after diligent assessment the code was deemed to be working as it should. This led to the conclusion that a periodic process in the classifier itself was causing this

noticeable behavior. The behavior was verified after we brute forced the subset of attack vectors using the method described in 2.2.

Even with the reduced color space the checkerboard pattern was clearly visible when analyzing the results from the brute force computations. Figure 5a shows an example image where the minimum confidence score is visualized for each pixel in the image from all the computed attack vectors. As can be seen in the image, the same checkerboard pattern is clearly visible.

The effect of even coordinates being more vulnerable to pixel modifications might be a side effect of the architecture that the targeted model uses. The model source code shows that the model uses convolutional blocks where convolutional layer stride is set to $(2, 2)$. This could cause the checkerboard pattern. If some filter kernel on a convolutional layer that has the stride of $(2, 2)$ is vulnerable to the pixel modification attack, then that effect would be duplicated to every other pixel while the kernel sweeps across the image dimensions while skipping every other coordinate.

3.3 Brute force confidence map result analysis

The brute force computations we performed for some handpicked images for both *mitosis-to-normal* and *normal-to-mitosis* images gave more information about the pixel positions for the successful attacks and a possible explanation why some of the attacks failed.

Figure 5a visualizes the minimum scores for each pixel that were computed for the attack vectors in the executed *mitosis-to-normal* brute force attack for this image. The original score for the image was 0.9874 and the lowest score that one of the pixel modifications achieved was 0.2689. The image shows that all the attack vectors successfully lowering the score in any meaningful way were situated in the middle of the dark spot in the image.

Figure 5b shows a similar mitosis image; however, in this image the dark spot is larger. This is an example of a failed *mitosis-to-normal* attack. The original score for this image was 0.99998 and the lowest score that any of the attack vectors achieved was 0.99979, so the best one-pixel change resulted in practically no change at all. Comparing this image to the successful *mitosis-to-normal* attack in Figure 5a shows that this time the pixel modifications that were in the middle of the dark spot had absolutely no effect at all, and the pixel modifications that had even the slightest effect to the score were the ones on the edge of the dark spot. This could indicate that the dark spot is so big that the one-pixel modification is not large enough change to fool the model.

Similar to the previously described *mitosis-to-normal* attacks, Figure 5c and Figure 5d show successful and failed *normal-to-mitosis* attacks. The successful attack in Figure 5c increased the score from original 0.09123 to 0.86350, but the failed attack in Figure 5d had practically no success at all with the original score of 4.29×10^{-7} and the highest achieved score of 1.04×10^{-6}. It seems that the successful *normal-to-mitosis* attacks require some kind of dark spots in the middle of the image that the attack pixel highlights by making the spot look bigger and this way fooling the model. If the image does not have a spot in the

middle of the image, then one pixel change is not enough to fool the model to think that there is a spot that would indicate a mitosis.

4 Conclusion

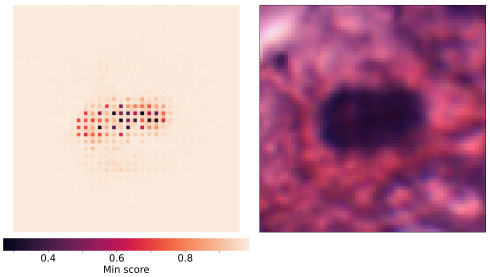
We have presented a way to systematically analyze the quality of one-pixel attacks. The target images were a set of digital pathology images and the target classifier tried to detect cancerous growth in them. We focused our efforts on the color and location of the attacks, as well as periodicity analysis through confidence maps. The tools we have used are able to reveal more information about the vulnerability of the classifier by pointing out the areas where successful attacks are more probable.

Chromatic analysis reveals that there are two clusters of attacks. It seems that the confidence score between the original and the adversarial images either stays low or, in the case of successful attacks, gets a rather big boost towards the wanted classification. Furthermore, the attack seems to be more effective the bigger the color difference is. As expected, this creates conflicting multi-objective optimization goals.

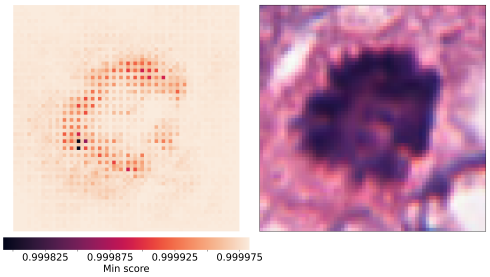
Spatial analysis reveals that the most sensitive areas for the attack are in the middle of the image. This is probably caused by the preprocessing, which produces images that have the prominent feature in the middle. This, in turn, causes the neural network classifier to focus on the middle of the image. Furthermore, combining the spatial and chromatic dimensions, pixels in successful attacks seem to appear inside the dark patches. Another common area is the edge of those dark patches. Taking into account the nature of the target images, this shows that color changes are prominent indications detected by the target model.

Periodicity analysis shows that some rows and columns are more susceptible to the attack. This stems from the features of the target classification model, which uses a neural network. It seems that a brute force mapping of classifier behavior is useful. The confidence maps illustrate that the most successful attacks are clustered around the dark middle areas of the images. It seems that it is difficult to realize a one-pixel attack if there is no clear dark area. This is caused by what the target classifier is trained to detect, and thus, focus on.

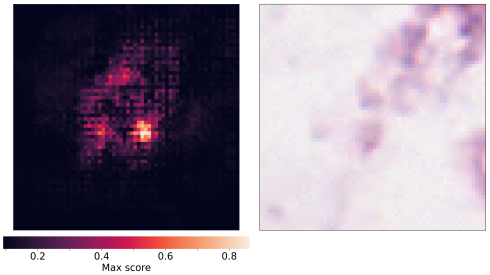
The methodology presented in this article is suitable for the analysis of any one-pixel attack, and not confined to the world of medical imaging. Our experiment used one dataset of such images. Therefore, the results may be skewed because of it and the target model used. Further experimentation could show the generalizability of the methods to other domains. The only requirement for the presented tools is to have access to a black-box classifier, which produces confidence scores. Such tools should be useful when assessing the quality of the classifier and its robustness. The need of including robustness metrics and mitigation methods to the toolbox of standard implementations seems like the correct direction in future research.



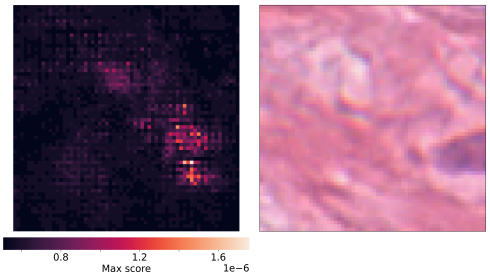
(a) Successful *mitosis-to-normal* attack



(b) Failed *mitosis-to-normal* attack



(c) Successful *normal-to-mitosis* attack



(d) Failed *normal-to-mitosis* attack

Fig. 5: Images showing examples of successful and failed *mitosis-to-normal* and *normal-to-mitosis* brute force attacks. In each subimage, on the right is the original image under brute force attack, and on the left is a heatmap visualization of maxscores or minscores array that are defined in algorithm 1. The heatmap image shows the pixel locations that had the biggest impact on the output score when the pixel color was changed.

Acknowledgments This work was funded by the Regional Council of Central Finland/Council of Tampere Region and European Regional Development Fund as part of the Health Care Cyber Range (HCCR) project of JAMK University of Applied Sciences Institute of Information Technology.

The authors would like to thank Ms. Tuula Kotikoski for proofreading the manuscript.

References

1. IBM code model asset exchange: Breast cancer mitosis detector. https://github.com/IBM/MAX-Breast-Cancer-Mitosis-Detector (2019)
2. Bharadwaj, H.K., Agarwal, A., Chamola, V., Lakkaniga, N.R., Hassija, V., Guizani, M., Sikdar, B.: A review on the role of machine learning in enabling iot based healthcare applications. IEEE Access **9**, 38859–38890 (2021). https://doi.org/10.1109/ACCESS.2021.3059858
3. Kamal, S.T., Hosny, K.M., Elgindy, T.M., Darwish, M.M., Fouda, M.M.: A new image encryption algorithm for grey and color medical images. IEEE Access **9**, 37855–37865 (2021). https://doi.org/10.1109/ACCESS.2021.3063237
4. Korpihalkola, J., Sipola, T., Kokkonen, T.: Color-optimized one-pixel attack against digital pathology images. In: Balandin, S., Koucheryavy, Y., Tyutina, T. (eds.) 2021 29th Conference of Open Innovations Association (FRUCT). vol. 29, pp. 206–213. IEEE (2021). https://doi.org/10.23919/FRUCT52173.2021.9435562
5. Korpihalkola, J., Sipola, T., Puuska, S., Kokkonen, T.: One-pixel attack deceives computer-assisted diagnosis of cancer. In: 2021 4th International Conference on Signal Processing and Machine Learning. pp. 100–106. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3483207.3483224
6. Latif, J., Xiao, C., Imran, A., Tu, S.: Medical imaging using machine learning and deep learning algorithms: A review. In: 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). pp. 1–5 (2019). https://doi.org/10.1109/ICOMET.2019.8673502
7. Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., Vasilakos, A.V.: Privacy and security issues in deep learning: A survey. IEEE Access **9**, 4566–4593 (2021). https://doi.org/10.1109/ACCESS.2020.3045078
8. Medical Image Analysis Group Eindhoven (IMAG/e): Tumor proliferation assessment challenge 2016. http://tupac.tue-image.nl/node/3 (2016)
9. Newaz, A.I., Haque, N.I., Sikder, A.K., Rahman, M.A., Uluagac, A.S.: Adversarial attacks to machine learning-based smart healthcare systems. In: GLOBECOM 2020 - 2020 IEEE Global Communications Conference. pp. 1–6 (2020). https://doi.org/10.1109/GLOBECOM42002.2020.9322472
10. Qayyum, A., Qadir, J., Bilal, M., Al-Fuqaha, A.: Secure and robust machine learning for healthcare: A survey. IEEE Reviews in Biomedical Engineering **14**, 156–180 (2021). https://doi.org/10.1109/RBME.2020.3013489
11. Sasubilli, S.M., Kumar, A., Dutt, V.: Machine learning implementation on medical domain to identify disease insights using tms. In: 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE). pp. 1–4 (2020). https://doi.org/10.1109/ICACCE49060.2020.9154960
12. Sipola, T., Kokkonen, T.: One-pixel attacks against medical imaging: A conceptual framework. In: Rocha, Á., Adeli, H., Dzemyda, G., Moreira, F., Ramalho Correia, A.M. (eds.) Trends and Applications in Information Systems and Technologies. Advances in Intelligent Systems and Computing, vol. 1365, pp. 197–203.

Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-72657-7_19
13. Sipola, T., Puuska, S., Kokkonen, T.: Model fooling attacks against medical imaging: A short survey. Information & Security: An International Journal (ISIJ) **46**, 215–224 (2020). https://doi.org/10.11610/isij.4615
14. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation **23**(5), 828–841 (2019). https://doi.org/10.1109/TEVC.2019.2890858
15. The European Union's Law Enforcement Agency, EUROPOL: How covid-19-related crime infected Europe during 2020. https://www.europol.europa.eu/sites/default/files/documents/how_covid-19-related_crime_infected_europe_during_2020.pdf (Nov 2020)
16. Vargas, D.V., Su, J.: Understanding the one-pixel attack: Propagation maps and locality analysis. In: Espinoza, H., McDermid, J., Huang, X., Castillo-Effen, M., Chen, X.C., Hernández-Orallo, J., Ó hÉigeartaigh, S., Mallah, R. (eds.) CEUR Workshop Proceedings. vol. 2640 (2020)
17. 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.: Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. Medical Image Analysis **54**, 111–121 (2019). https://doi.org/10.1016/j.media.2019.02.012
18. Wang, W., Sun, J., Wang, G.: Visualizing one pixel attack using adversarial maps. In: 2020 Chinese Automation Congress (CAC). pp. 924–929. IEEE (2020). https://doi.org/10.1109/CAC51589.2020.9327603
19. Watson, M., Al Moubayed, N.: Attack-agnostic adversarial detection on medical data using explainable machine learning. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 8180–8187 (2021). https://doi.org/10.1109/ICPR48806.2021.9412560
20. Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE **109**(5), 820–838 (2021). https://doi.org/10.1109/JPROC.2021.3054390