Model Fooling Threats Against Medical Imaging

Tuomo Sipola(✉), Tero Kokkonen, and Mika Karjalainen

Institute of Information Technology,
JAMK University of Applied Sciences,
Jyväskylä, Finland
{tuomo.sipola, tero.kokkonen, mika.karjalainen}@jamk.fi

Abstract. Automatic medical image diagnosis tools are vulnerable to modern model fooling technologies. Because medical imaging is a way of determining the health status of a person, the threats could have grave consequences. These threats are not only dangerous to the individual but also threaten the trust in modern diagnosis methods and in the healthcare sector as a whole. As newer diagnosis tools are based on artificial intelligence and machine learning, they can be exploited using attack technologies such as image perturbations, adversarial patches, adversarial images, one-pixel attacks, and training process tampering. These methods take advantage of the non-robust nature of many machine learning models created to solve medical imaging classification problems, such as determining the probability of cancerous cell growth in tissue samples. In this study, we review the current state of these attacks and discuss their effect on medical imaging. By comparing the known attack methods and their use against medical imaging, we conclude with an evaluation of their possible use against medical imaging.

Keywords: Deep learning · Cyber security · Adversarial examples · Medical Imaging

1 Introduction

The goal of the research is to examine the literature related to potential model fooling attacks against medical imaging, with digital pathology as the main interest. This article is an extension of our short survey originally presented in the Second International Scientific Conference *"Digital Transformation, Cyber Security and Resilience"* (DIGILIENCE 2020) and published in the conference-related special issues of Information & Security: An International Journal [48]. This research is expanded from the original as follows: we have identified more essential publications related to the topic and presented them in a new manner. In addition, we have restructured this paper to better reflect the contents of the identified research literature.

In the modern digitalised world, Artificial Intelligence (AI) based solutions are utilised extensively in everyday life. For example, paper [38] introduces an AI-based healthcare assistant. Heart functioning is analysed and predicted with neural networks by using electrocardiogram (ECG) data in the studies [40,41]

This is a pre-print of the following chapter: Tuomo Sipola, Tero Kokkonen, and Mika Karjalainen, Model Fooling Threats Against Medical Imaging, published in Artificial Intelligence and Cybersecurity: Theory and Applications, edited by Tuomo Sipola, Tero Kokkonen, and Mika Karjalainen, 2022, Springer. Reproduced with permission of Springer. The final authenticated version is available online at: https://dx.doi.org/10.1007/978-3-031-15030-2_13

and similarly, electroencephalogram (EEG) data is analysed by AI for detecting brain tumors [33]. Syam and Marapareddy used deep neural networks for network intrusion detection, heart disease prediction and for skin cancer classification [52].

The usage of sub-disciplines of AI, Machine learning (ML) and Deep Learning (DL) based solutions is rapidly increasing in the medical imaging for prediction and decision making by itemizing and labeling disease patterns from image samples [25]. The large amounts of information available makes the medical domain very interesting for researchers so that new applications can be developed [45]. The tremendous development of medical imaging has produced advances in diagnostics and prediction of diseases [13,3]. The benefit achieved by the DL in the analysis of modern medical big data is the capability for algorithmic realisation of the various associations and capability to combine learned lines or edges of low level to the higher-level shapes [21].

The vast development of machine learning has produced several modern examples of applying ML/DL for the medical imaging as computer-aided diagnosis (CAD) tools. Comprehensive review for ML in medicine is presented by authors of paper [39]. Hussein et al. studied lung and pancreatic tumor characterization with DL whereas Lu et al. utilised ensemble learning with data mining for predicting recurrent ovarian cancer. Among others, during this year, utilisation of ML/DL for medical image classification and detection is studied for example with brain tumors in [43,49,42] and breast cancer in [11,44,37]. It should also be noticed that developing AI for healthcare is a highly technical subject but in addition with usage of AI for healthcare there are ethical, legal and social challenges involved such as 'Data ownership, confidentiality and consent' or 'Medical moral and professional responsibility' [9].

Modern networked and digitalized cyber domain is an extremely complex entity that comprises unpredictable phenomena. A classical example of that complexity is a cyber attack against an electricity company, which may endanger the patient safety of the hospital. Finland's cyber security strategy [46] classifies healthcare as an area that is vulnerable to cyber security issues, and states that these issues will be more important in the future. As known, there are several cyber attacks executed globally against healthcare infrastructure, and healthcare infrastructure is seen as valuable target for cyber attacks or an intrusion. The International Criminal Police Organization (INTERPOL) states that cyber attacks' target is shifting towards governments and critical health infrastructure during the ongoing COVID-19 pandemic [20].

As can be seen, ML/DL applications are widely applied in the medical imaging and simultaneously, the overall medical cyber domain is realised as a potential target for the cyber attacks. In that sense, our study focuses on the model fooling threats against medical imaging. During this study following threat categories are identified: (i) adversarial images, (ii) adversarial patches, (iii) one-pixel attacks and (iv) training process tampering.

We investigated literature related to the possible cybersecurity threat vectors using using a scoping review method. According to Munn et al. scoping review is a suitable method for the search for scientific gaps in the research area, or building the knowledge base or the synthesis of literature to confirm the research results [32]. In this paper, the point of scoping review is to seek support from previous research for the findings of this research, thus building a stronger knowledge base for the phenomenon. In the scoping review, we used Google Scholar and IEEE databases. Searches were performed by using the following search parameters: fooling neural networks, adversarial attack / adversarial example and medical imaging. Studies in English related to the medical domain were selected. Furthermore, studies with actual application of the attacks were included. In addition, some essential methodology studies are mentioned.

The paper is organised as follows: Neural network fooling and the specific categories of attacks are discussed in section 2 and its subsections. The research is concluded with the found future research topics in section 4.

2 Fooling Deep Neural Networks in Medical Imaging

Deep learning tries to combine simple concepts into a representation of the actual object. This happens by creating an artificial neural network of interconnected nodes [17]. The complex nature of these networks makes them susceptible to unexpected attacks, which make the network to output completely reverse results that are unlike the expected outcome. A reverse result in medical imaging could be harmful to the patient.

Adversarial attacks against deep learning image classifiers are plentiful. In a white-box attack, the attacker knows the internal workings of the classifiers. This is usually useful when using the neural network gradient as a way of finding adversarial examples. On the other hand, black-box attacks are performed against a system that has only its image input and classification result exposed to the attacker. The attack methods use optimization to find examples that produce the most diverging classification scores [56]. Computer vision is especially affected by these threats because deep neural networks are the most prominent method. Akhtar and Mian estimate that the Carlini & Wagner [8] and Universal perturbations [30] are the strongest methods. Both are white-box attacks, so they need the complete knowledge of the inner workings of the target classifier [2]. Furthermore, Afifi et al. demonstrate that simple color constancy errors can change the classification of a natural image [1]. Another kind of proof of the power of adversarial images is that they can fool human experts. Chuquicusma et al. have shown that images produced by generative adversarial networks can fool radiologists [10].

Since many methods use gradient as the guiding principle for the optimization, gradient masking and obfuscation could help to defend against these attacks. This would mislead the attacks or make the attack optimization very difficult to achieve. Another defence is the use of robust optimization. Robust classifiers are less likely to behave in an unexpected manner, such as falling for an adversarial image. This could be achieved, e.g., with adversarial retraining. The third defence could be adversarial example detection before the input images are fed to the real classifier [56,27]. Tizhoosh and Pantanowitz mention adversarial attacks as one of the challenges facing digital pathology. They raise the question whether minimal artifacts could reduce the reliability of neural network classifiers. This might be caused by the old problem of overfitting in artificial intelligence [54]. Akhtar and Mian propose three ways of defending against adversarial attacks. Firstly, modified training during learning or modified input during testing can be used. Secondly, they suggested modifying deep neural networks and their architecture. Thirdly, for unseen examples, an external model could be used to act as a network add-on [2]. The point of intervention and defence against these attacks is also a problem to be solved, which will probably need regulatory best practices since the problem resembles that of trying to counteract ever developing hacking attempts [14]. A recent survey by Apostolidis and Papakostas on adversarial attacks against medical image analysis discusses the robustness of deep neural networks. It identifies many image modalities that have been attacked: X-ray images, magnetic resonance imaging (MRI), computer tomography scans (CT), retinal images, histology and skin. In addition to the modalities, the survey lists attacks, their target models, detection methods and defences. The authors emphasize the need for robust models in automated medical imaging [4].

However, not all applications of adversarial methods are malicious. Generative adversarial networks can also be used for synthesizing data samples [29]. Another application is to use generative adversarial methods to inpaint medical images that contain areas of missing data [5].

3 Attack Types

Based on the literature introduced in this study, Figure 1 shows the most obvious attack vectors against medical imaging neural networks. The two proposed attack vectors are changing the training process to create a faulty AI model and modifying the input images, so that the classification goes wrong even with a correctly working AI model.

There are several ways to attack against medical imaging. The main methods can be categorized as (i) adversarial images, (ii) adversarial patches, (iii) one-pixel attack and (iv) training process tampering. Table 1 shows the identified attack methods and their use against medical imaging. The following subsections discuss each of these methods in more detail, introducing the methods themselves, and discussing their applications.

3.1 Adversarial Images

Adversarial images are images that are somehow changed by adding perturbation to create an image that is misclassified. As shown by Nguyen et al., it is possible to produce images that are unrecognizable to humans, but that are classified with 99.99% confidence by deep neural networks. Firstly, their adversarial examples include pictures that resemble noise generated by an evolutionary algorithm

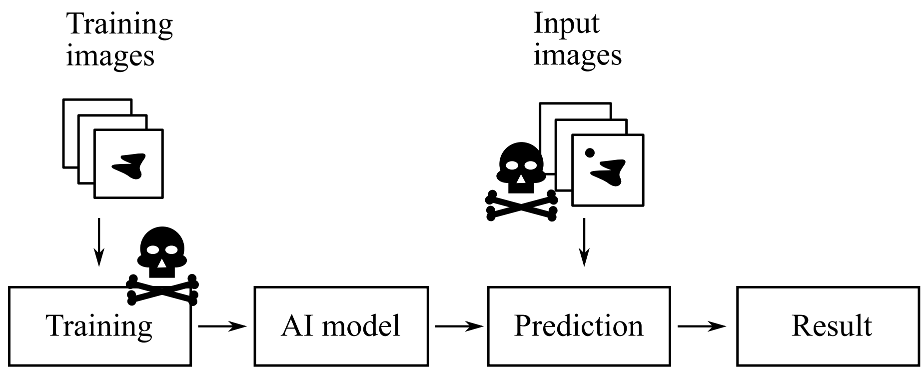

Fig. 1. The most prominent attack vectors described in literature. Tampering with training compromises the automated detection pipeline from the beginning. Modifying input images is perhaps the easier attack and compromises the results of automated detection.

Table 1. Adversarial methods against artificial neural networks, and their implementations in the medical domain.

Method	References	Medical domain
Adversarial images	[34], [31], [6], [28], [12], [19]	[35], [53], [15], [29], [5], [26]
Adversarial patches	[7]	[15]
One-pixel attack	[50], [51], [24], [55]	[36], [47], [23], [22]
Training tampering	[18], [57]	

using direct encoding. Secondly, their other adversarial examples resemble wave patterns and lattices, which have been created by an evolutionary algorithm using indirect encoding. Their evolutionary optimization uses the classifying deep neural network as the fitness function, which makes the approach a black-box method. [34]

Moossavi-Dezfooli et al. present the DeepFool algorithm that finds perturbations to deceive deep neural networks. They use a gradient descent algorithm to find those perturbations. The combination of an image and the perturbation is falsely classified as representing something that it does not. [31] Athalye et al. raise the question that viewpoint shifts, camera noise, and transformations can make adversarial examples less effective. They created a 3D-printed turtle that is classified as rifle from images taken of it in the physical world. The optimization process takes into account the expectation of transformation, which creates more robust adversarial examples. [6]

Some other examples of adversarial images include those generated using adversarial noise [28], using a generative approach to fool black-box classifiers [12] and gradient shielding to identify sensitive regions where attacks could be executed [19].

Medical images have been used as targets for these kinds of adversarial images. Paschali et al. studied neural network performance under extreme inputs such as noise, outliers, and ambiguous data. They used fast gradient sign, DeepFool and saliency map attacks to create the adversarial images. They performed the attacks on skin lesion images and whole brain imaging [35]. Taghanaki et al. used three types of adversarial attacks: gradient-based, score-based and decision-based. These added perturbations to X-ray images producing images that look quite natural in some cases [53]. Finlayson et al. used projected gradient descent to create visually unnoticeable perturbations against fundoscopy, chest X-ray, and dermoscopy images [15].

Ma et al. created adversarial images in medical imaging domain using unnoticeable perturbations. They go on to claim that medical images can be more vulnerable than natural images in this context. Firstly, they suggest that medical images have larger high attention regions, which draw unnecessary attention from the neural network. Secondly, modern neural networks are designed for natural images, causing them to overparametrize for medical images. Furthermore, a simple adversarial image detector classifier is sufficient to protect the actual classifier from most of the attacks. [26].

3.2 Adversarial Patches

Adversarial patches can be applied onto images to output any target class. These patches can be natural, meaning a cut-and-pasted part of an existing image, or generated using optimization, resulting in wild-looking but successful patches when applied. According to Brown et al., even small patches can shift the focus of the classifier to the patch and change the classification of the scene. Suitable patches are found with similar optimization as with adversarial images. [7]

There have been examples of adversarial patches used against medical imaging. Finlayson et al. demonstrated that this method works against fundoscopy, chest X-ray, and dermoscopy images. Furthermore, they tested natural patches, patches built on the victim model and patches built on another independent model later used as attacks against the victim model. [15]

3.3 One-pixel Attacks

One-pixel attack means that the alteration of color values of a single pixel will cause misclassification. Su et al. have shown that this extremely limited attack is successful against natural images. They used differential evolution optimization against the black-box classifier to find successful one-pixel examples [50]. Furthermore, they propose a variation of the attack with multiple objectives [51]. Gilmer et al. propose that small perturbations are adversarial against machine learning models because of high-dimensional geometry of the data manifold. [16]

Kügler et al. created simple problems about pose estimation of surgical tools in order to localize areas where one-pixel attacks were lucrative. They discovered that the vulnerable areas of the image are close to the decision boundary. [24]

Vargas et al. propose propagation maps to illustrate how much the perturbations affect neural network layers. They discovered that complex neural networks let the single pixel propagate widely causing it to create unreasonable consequences to the classification result. Attacks against pixels near a successful one are also quite effective. [55]

There are not many examples of one-pixel attacks against real medical imaging data. Paul et al. attacked against the National Lung Screening Trial (NLST) dataset using a one-pixel attack. They also used fast gradient signed method (FGSM) attack, which was more successful. They applied an ensemble defence strategy to create more robust classifiers [36]. The concept of using one-pixel attacks against whole slide images was explored by Sipola and Kokkonen [47], and implemented by Korpihalkola et al. using an existing database of those images [23]. The attack was refined by optimizing the color so that the adversarial pixels would be less prominent to the human eye [22].

3.4 Training Process Tampering

A backdoored neural network has been trained with malicious training material that cause it to react in unexpected ways when given specific input. The act of infiltrating training data with malicious samples is called poisoning. Yang et al. used direct gradient method and auto-encoders to generate poisoned data for neural network training. [57] Gu et al. present this idea of including a hidden backdoor detector inside the classifier by using crafted training data. They demonstrate this threat using traffic signs, which causes the classifier to detect a stop sign as a speed limit sign. [18]

4 Conclusion

Machine learning based solutions are enormously and successfully used in healthcare, especially in the medical imaging for prediction and decision making of possible tumors. Medical imaging and analysis methods related to medical imaging are not safe from model fooling attacks. Suitable research exploits have been shown to successfully fool neural network models in this domain. The most prominent exploits are (i) adversarial images, (ii) adversarial patches, (iii) one-pixel attacks, and (iv) training process tampering. These main types of attacks against medical imaging are present in the scientific studies included in this survey. Based on the conducted scoping review, future research could include a comprehensive systematic literature review of the phenomenon, especially for specific imaging modalities or attack methods. Further investigation needs to be focused on the deep neural network methods used in medical classifiers. The underlying causes and robustness of those networks are not yet clear and the theoretical considerations still unresolved.

Acknowledgments. This research is partially 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 and The Cyber Security Network of Competence Centres for Europe (CyberSec4Europe) project of the Horizon 2020 SU-ICT-03-2018 program. The authors would like to thank Ms. Tuula Kotikoski for proofreading the manuscript.

References

1. Afifi, M., Brown, M.S.: What else can fool deep learning? Addressing color constancy errors on deep neural network performance. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 243–252 (2019)
2. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access **6**, 14410–14430 (2018). https://doi.org/10.1109/ACCESS.2018.2807385
3. Al-Sharify, Z.T., Al-Sharify, T.A., Al-Sharify, N.T., naser, H.Y.: A critical review on medical imaging techniques (CT and PET scans) in the medical field. IOP Conference Series: Materials Science and Engineering **870**, 012043 (jul 2020). https://doi.org/10.1088/1757-899x/870/1/012043
4. Apostolidis, K.D., Papakostas, G.A.: A survey on adversarial deep learning robustness in medical image analysis. Electronics **10**(17), 2132 (Sep 2021). https://doi.org/10.3390/electronics10172132, http://dx.doi.org/10.3390/electronics10172132
5. Armanious, K., Mecky, Y., Gatidis, S., Yang, B.: Adversarial inpainting of medical image modalities. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3267–3271 (2019). https://doi.org/10.1109/ICASSP.2019.8682677
6. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: Proceedings of the 35th International Conference on Machine Learning. vol. 80, pp. 284–293. PMLR, Stockholm, Sweden (2018)

7. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665v2 (2018)
8. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 39–57 (2017). https://doi.org/10.1109/SP.2017.49
9. Carter, S.M., Rogers, W., Win, K.T., Frazer, H., Richards, B., Houssami, N.: The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. The Breast **49**, 25–32 (2020). https://doi.org/10.1016/j.breast.2019.10.001
10. Chuquicusma, M.J.M., Hussein, S., Burt, J., Bagci, U.: How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 240–244 (2018). https://doi.org/10.1109/ISBI.2018.8363564
11. Cristovao, F., Cascianelli, S., Canakoglu, A., Carman, M., Nanni, L., Pinoli, P., Masseroli, M.: Investigating deep learning based breast cancer subtyping using pan-cancer and multi-omic data. IEEE/ACM Transactions on Computational Biology and Bioinformatics pp. 1–1 (2020). https://doi.org/10.1109/TCBB.2020.3042309
12. Deng, Y., Zhang, C., Wang, X.: A multi-objective examples generation approach to fool the deep neural networks in the black-box scenario. In: 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC). pp. 92–99 (2019). https://doi.org/10.1109/DSC.2019.00022
13. Doi, K.: Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics **31**(4–5), 198–211 (2007)
14. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
15. Finlayson, S.G., Chung, H.W., Kohane, I.S., Beam, A.L.: Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:1804.05296v3 (2019)
16. Gilmer, J., Metz, L., Faghri, F., Schoenholz, S.S., Raghu, M., Wattenberg, M., Goodfellow, I., Brain, G.: The relationship between high-dimensional geometry and adversarial examples. arXiv preprint arXiv:1801.02774 (2018)
17. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org
18. Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S.: BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. IEEE Access **7**, 47230–47244 (2019). https://doi.org/10.1109/ACCESS.2019.2909068
19. Gu, Z., Hu, W., Zhang, C., Lu, H., Yin, L., Wang, L.: Gradient shielding: Towards understanding vulnerability of deep neural networks. IEEE Transactions on Network Science and Engineering pp. 1–1 (2020). https://doi.org/10.1109/TNSE.2020.2996738
20. INTERPOL, The International Criminal Police Organization: INTERPOL report shows alarming rate of cyberattacks during COVID-19 (Aug 2020), https://www.interpol.int/en/News-and-Events/News/2020/INTERPOL-report-shows-alarming-rate-of-cyberattacks-during-COVID-19, accessed: 4 December 2020
21. Ker, J., Wang, L., Rao, J., Lim, T.: Deep learning applications in medical image analysis. IEEE Access **6**, 9375–9389 (2018). https://doi.org/10.1109/ACCESS.2017.2788044

22. 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
23. Korpihalkola, J., Sipola, T., Puuska, S., Kokkonen, T.: One-pixel attack deceives computer-assisted diagnosis of cancer. In: Proceedings of the 4th International Conference on Signal Processing and Machine Learning (SPML 2021), August 18–20, 2021, Beijing, China. ACM, New York, USA (2021). https://doi.org/10.1145/3483207.3483224
24. Kügler, D., Distergoft, A., Kuijper, A., Mukhopadhyay, A.: Exploring adversarial examples. In: Stoyanov, D., Taylor, Z., Kia, S.M., Oguz, I., Reyes, M., Martel, A., Maier-Hein, L., Marquand, A.F., Duchesnay, E., Löfstedt, T., Landman, B., Cardoso, M.J., Silva, C.A., Pereira, S., Meier, R. (eds.) Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Lecture Notes in Computer Science, vol. 11038, pp. 70–78. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-02628-8_8
25. 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
26. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition **110**, 107332 (2020). https://doi.org/10.1016/j.patcog.2020.107332
27. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv preprint arXiv:1706.06083v4 arXiv:1706.06083 (Sep 2019)
28. Mihajlović, M., Popović, N.: Fooling a neural network with common adversarial noise. In: 2018 19th IEEE Mediterranean Electrotechnical Conference (MELECON). pp. 293–296 (2018). https://doi.org/10.1109/MELCON.2018.8379110
29. Mikołajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW). pp. 117–122 (2018). https://doi.org/10.1109/IIPHDW.2018.8388338
30. Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 86–94. IEEE Computer Society, Los Alamitos, CA, USA (jul 2017). https://doi.org/10.1109/CVPR.2017.17
31. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2574–2582 (2016). https://doi.org/10.1109/CVPR.2016.282
32. Munn, Z., Peters, M.D.J., Stern, C., Tufanaru, C., McArthur, A., Aromataris, E.: Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. BMC Medical Research Methodology **18**, 143 (2018). https://doi.org/10.1186/s12874-018-0611-x
33. Murugesan, M., Sukanesh, R.: Automated Detection of Brain Tumor in EEG Signals Using Artificial Neural Networks. In: 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies. pp. 284–288 (2009). https://doi.org/10.1109/ACT.2009.77

34. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 427–436 (2015). https://doi.org/10.1109/CVPR.2015.7298640
35. Paschali, M., Conjeti, S., Navarro, F., Navab, N.: Generalizability vs. robustness: adversarial examples for medical imaging. arXiv preprint arXiv:1804.00504 (2018)
36. Paul, R., Schabath, M., Gillies, R., Hall, L., Goldgof, D.: Mitigating adversarial attacks on medical image understanding systems. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 1517–1521 (2020). https://doi.org/10.1109/ISBI45749.2020.9098740
37. Poonguzhali, N., Dharani, V., Nivedha, R., Ruby, L.S.: Prediction of breast cancer using electronic health record. In: 2020 International Conference on System, Computation, Automation and Networking (ICSCAN). pp. 1–6 (2020). https://doi.org/10.1109/ICSCAN49426.2020.9262398
38. Rai, S., Raut, A., Savaliya, A., Shankarmani, R.: Darwin: Convolutional neural network based intelligent health assistant. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). pp. 1367–1371 (2018). https://doi.org/10.1109/ICECA.2018.8474861
39. Rajkomar, A., Dean, J., Kohane, I.: Machine learning in medicine. The New England Journal of Medicine **380**(14), 1347–1358 (2019). https://doi.org/10.1056/NEJMra1814259
40. Rastgar-Jazi, M., Fernando, X.: Detection of heart abnormalities via artificial neural network: An application of self learning algorithms. In: 2017 IEEE Canada International Humanitarian Technology Conference (IHTC). pp. 66–69 (2017). https://doi.org/10.1109/IHTC.2017.8058202
41. Ravish, D.K., Shanthi, K.J., Shenoy, N.R., Nisargh, S.: Heart function monitoring, prediction and prevention of heart attacks: Using artificial neural networks. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I). pp. 1–6 (2014). https://doi.org/10.1109/IC3I.2014.7019580
42. Razzaq, S., Mubeen, N., Kiran, U., Asghar, M.A., Fawad, F.: Brain tumor detection from mri images using bag of features and deep neural network. In: 2020 International Symposium on Recent Advances in Electrical Engineering Computer Sciences (RAEE CS). vol. 5, pp. 1–6 (2020). https://doi.org/10.1109/RAEECS50817.2020.9265768
43. Ruiz, L., Martín, A., Urbanos, G., Villanueva, M., Sancho, J., Rosa, G., Villa, M., Chavarrías, M., Pérez, Á., Juarez, E., Lagares, A., Sanz, C.: Multiclass brain tumor classification using hyperspectral imaging and supervised machine learning. In: 2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS). pp. 1–6 (2020). https://doi.org/10.1109/DCIS51330.2020.9268650
44. Salama, W.M., Elbagoury, A.M., Aly, M.H.: Novel breast cancer classification framework based on deep learning. IET Image Processing **14**(13), 3254–3259 (2020). https://doi.org/10.1049/iet-ipr.2020.0122
45. 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
46. Secretariat of the Security Committee: Finland's Cyber security Strategy, Government Resolution 3.10.2019. https://turvallisuuskomitea.fi/wp-content/uploads/2019/10/Kyberturvallisuusstrategia_A4_ENG_WEB_031019.pdf (Oct 2019)

47. 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. (eds.) Trends and Applications in Information Systems and Technologies. WorldCIST 2021. Advances in Intelligent Systems and Computing, vol. 1365, pp. 197–203. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72657-7_19
48. Sipola, T., Puuska, S., Kokkonen, T.: Model fooling attacks against medical imaging: A short survey. Information & Security: An International Journal **46**(2), 215–224 (2020). https://doi.org/10.11610/isij.4615
49. Someswararao, C., Shankar, R.S., Appaji, S.V., Gupta, V.: Brain tumor detection model from mr images using convolutional neural network. In: 2020 International Conference on System, Computation, Automation and Networking (ICSCAN). pp. 1–4 (2020). https://doi.org/10.1109/ICSCAN49426.2020.9262373
50. 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
51. Su, J., Vargas, D.V., Sakurai, K.: Attacking convolutional neural network using differential evolution. IPSJ Transactions on Computer Vision and Applications **11**(1), 1–16 (2019)
52. Syam, R., Marapareddy, R.: Application of deep neural networks in the field of information security and healthcare. In: 2019 SoutheastCon. pp. 1–5 (2019). https://doi.org/10.1109/SoutheastCon42311.2019.9020553
53. Taghanaki, S.A., Das, A., Hamarneh, G.: Vulnerability Analysis of Chest X-Ray Image Classification Against Adversarial Attacks. In: Stoyanov, D., Taylor, Z., Kia, S.M., Oguz, I., Reyes, M., Martel, A., Maier-Hein, L., Marquand, A.F., Duchesnay, E., Löfstedt, T., Landman, B., Cardoso, M.J., Silva, C.A., Pereira, S., Meier, R. (eds.) Understanding and Interpreting Machine Learning in Medical Image Computing Applications. pp. 87–94. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-02628-8_10
54. Tizhoosh, H.R., Pantanowitz, L.: Artificial intelligence and digital pathology: challenges and opportunities. Journal of pathology informatics **9** (2018)
55. Vargas, D.V., Su, J.: Understanding the one-pixel attack: Propagation maps and locality analysis. arXiv preprint arXiv:1902.02947 (2019)
56. Xu, H., Ma, Y., Liu, H.C., Deb, D., Liu, H., Tang, J.L., Jain, A.K.: Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing **17**(2), 151–178 (2020). https://doi.org/10.1007/s11633-019-1211-x
57. Yang, C., Wu, Q., Li, H., Chen, Y.: Generative poisoning attack method against neural networks. arXiv preprint arXiv:1703.01340 (2017)