ISSN 0146-4116, Automatic Control and Computer Sciences, 2022, Vol. 56, No. 8, pp. 934–941. © Allerton Press, Inc., 2022.
Russian Text © The Author(s), 2022, published in Problemy Informatsionnoi Bezopasnosti, Komp'yuternye Sistemy.

Adversarial Machine Learning Protection Using the Example of Evasion Attacks on Medical Images

E. A. Rudnitskaya[a] and M. A. Poltavtseva[a, *]

[a] *Peter the Great St. Petersburg Polytechnic University, St. Petersburg, 195251 Russia*

**e-mail: poltavtseva@ibks.spbstu.ru*

Received April 11, 2022; revised April 18, 2022; accepted May 23, 2022

Abstract—This work considers evasion attacks on machine learning (ML) systems that use medical images in their analysis. Their systematization and a practical assessment of feasibility are carried out. Existing protection techniques against ML evasion attacks are presented and analyzed. The features of medical images are given and the formulation of the problem of evasion attack protection for these images based on several protective methods is provided. The authors have identified, implemented, and tested the most relevant protection methods on practical examples: an analysis of images of patients with COVID-19.

Keywords: AML attacks, protection of machine learning systems, evasion attacks, adversarial attacks, medical images, machine learning

DOI: 10.3103/S0146411622080211

Machine learning (ML) has become an integral part of modern life, starting from the medical field (medical diagnoses with subsequent treatment) and ending with finding friends in social networks. In recent years there has been an increase in the amount of available data and computing power, which as a result has led to a wider use of ML methods as a convenient tool for data classification and clustering and other tasks in the medical field. The relevance of the study comes from the increase in the number of different attacks on medical ML systems. In this regard, it is necessary to analyze existing vulnerabilities and improve known protection methods.

Medical images for diagnosing and detecting diseases are black-and-white CT images. For example, using such images, the following diseases can be diagnosed: Parkinson's disease, breast cancer, lung cancer, brain cancer, pneumonia, coronavirus infection, and others.

In [1], an attack is considered in the context of medical systems (medical image recognition system), where the authors note that evasion attacks on medical images can be more successful than attacks on natural (ordinary) images. That is, a successful attack requires less perturbation, and black-and-white computed tomography images are used to diagnose diseases, the noise pollution of which may not be noticed by the human eye, which excludes the possibility of the sample being put aside by a specialist.

Source [2] considers a study by Israeli infosecurity specialists who managed to attack a medical system that diagnoses lung cancer. They managed to edit the CT images of patients, and neither doctors nor the programs they use could recognize the fake. In most cases, the diagnoses were wrong.

The researchers noted that such an attack can have a great impact on people's lives, affecting both the health and moral order of individuals, and even influencing policy.

1. AML ATTACKS

1.1. Review of the Literature

Works [3, 4] consider the classification of adversarial machine learning (AML) attacks. According to sources, three types of AML attacks can be distinguished: poisoning, garbling, and exploratory. Knowledge of the intruder is also taken into account: they may know the learning algorithm or the data involved in learning, or both, or nothing at all. Thus, three attack strategies can be distinguished: white box, gray box, and black box. According to [5, 6], attacks can be classified according to the method of impact, security breach, and specificity. The disadvantage of all existing AML-attack systematizations

Table 1. Systematization of AML attacks

Security breach	Object of attack	Evasion attacks	Poisoning attacks	Exploratory attacks
Accessibility violation	–	–	–	–
Integrity breach	Training set	–	+ PO	–
	Input data	+ PP	–	–
Privacy violation	Training set	–	–	+ PP
	Model	–	–	+ PP

is the lack of a unified and complete systematization of the most important criteria, as well as security violations regarding the so-called component parts of the system (such as input data, the system itself, and the output result).

In [1], an attack is considered in the context of medical systems (medical image recognition system), where the authors note that evasion attacks on medical images can be more successful than attacks on natural (ordinary) images. That is, a lesser perturbation is required for a successful attack. In [7], an adversarial attack on medical ML systems is considered in the context of insurance payments, when this type of attack is used for profit.

In [8], the authors test two modern neural networks for resistance to ten different types of evasion attacks, demonstrating that these models are indeed vulnerable to this type of attack. In [9], the neural network InceptionV3 is also tested using chest X-ray images and histological images of malignant tumors. The authors determined that neural networks trained to classify histological images were more resistant to attacks than networks trained to classify chest CT images.

Work [10] considers a universal adversarial attack on deep neural networks for the classification of medical images. The authors found that deep neural networks are vulnerable to both untargeted attacks of this type and targeted ones. It was also found that the vulnerability to this attack depended very little on the model architecture. The authors emphasize that adversarial learning, which is known to be an effective method of adversarial protection, increases the resilience of deep neural networks to a universal adversarial attack only in very few cases.

The disadvantage of the reviewed studies is the lack of consideration of various approaches to protection against evasion attacks in the context of the medical field for solving a specific problem—for example, for classifying CT images. This problem is considered in this work.

1.2. Systematization of AML Attacks

Based on already existing classifications [3–6], the following systematization of AML attacks was compiled, shown in Table 1. The authors identified two modes of operation of the protected object (ML system): operation mode and training mode. The classification is based on the type of security violation or violation of the availability, integrity, and confidentiality of objects (from which the training set, input data, and model are selected).

Next, an evasion attack and methods of protection against it will be considered, since this type of attack, according to the authors, is the most dangerous. A successful evasion attack can affect medical actions, and thus the lives and health of people (compared to exploratory attacks) in the context of the medical field. This method does not require access to the training sample, like poisoning attacks, so it is applicable to systems that were created in a trusted execution environment at the stage of their operation.

1.3. Features of Medical Image Analysis Systems As an Object of Attack

In medicine, ML algorithms can be applied in various fields. Hereinafter, the field of diagnostics, in which medical images are directly applied, will be considered. Medical images in the diagnosis and detection of diseases are often black-and-white CT ones. For example, using such images, the following dis-

Table 2. Evasion attack protection methods

Method class	Defense mechanism	Object of assessment/impact	Known limitations
Proactive methods	Including adversarial sampling in training	Training set	Does not always cover the full range of possible adversarial samples
	Random data changing	Training sample	Potential failure when attacking the white box model
	Random data noising	Training sample	Potential failure when attacking the white box model. Decreased model performance
Active methods	Automatic noise reduction	Input data	Potential failure when attacking the white box model

eases can be diagnosed: Parkinson's disease, breast cancer, lung cancer, brain cancer, pneumonia, coronavirus, and others.

In [1], the authors note that evasion attacks on medical images can be more successful than attacks on natural (ordinary) images. This is due to several reasons:

(A) Some medical images have complex biological textures resulting in higher gradient regions that are sensitive to small opposing perturbations;

(B) Modern neural networks that are part of ML systems and designed for large-scale natural image processing can be overparameterized for medical imaging problems (too complex for relatively simple classification problems), which leads to a high vulnerability to attacks (a small perturbation gives a "strong" reaction; strong influence of the loss function on the final result).

It can also be noted that black-and-white CT images are often used for the diagnosis of diseases, the noise pollution of which may not be noticed by the human eye, which excludes the possibility of the sample being put aside by a specialist.

2. EVASION ATTACKS PROTECTION METHODS

All evasion attacks protection methods can be divided into two main categories: proactive and active methods [11]. Proactive methods warn of a possible attack, while active ones fight in real time and try to eliminate or mitigate the possibility of an attack. Let us give a brief description of known solutions.

An adversarial learning method [12, 13] is an attempt to improve the reliability of a neural network by training it with adversarial samples. The disadvantages of this method are the complexity and inability to cover the full range of possible competitive samples in all cases.

A randomization method, which includes random noise pollution and/or data modification [14–16], is based on a random change in input data: for example, image size, image filling with zeros, image compression, and bit depth reduction before being fed to the classifier. This protection attempts to transform hostile effects into random effects, which are not a problem for most neural networks. If the intruder knows the system construction (white box), the protection can be compromised, taking into account the possibility of its existence. The disadvantage is also a decrease in the model performance.

A compression method [17] is an active protection method. The fact that a sample is adversarial is determined by comparing the model's predictions on the original and compressed images. Data cleaning methods (GAN-based methods [18, 19]) are aimed at removing perturbation from a sample recognized as adversarial. However, this protection may not be effective for the attack described by Carlini and Wagner in [20].

An automatic noise suppression method [21] is similar to the previous one: the automatic encoder learns the variety of benign samples, and the detector distinguishes between hostile and benign samples based on the learned sample. The reformer, on the other hand, transforms a malicious sample into a legitimate one. The algorithm in [22] uses a loss function to minimize the difference at the feature level between benign and hostile samples. The disadvantages of this method are the relative complexity in model building, and this method may not always be effective against white box attacks.

Table 2 provides a systematization of methods for protecting against evasion attacks (or adversarial attacks).

Table 3. Number of training and test samples

Data set	Number of images from the COVID class	Number of images from the NORMAL class	Total number of images
Training samples	1700	1841	3541
Test samples	931	898	1829

Table 4. Neural network training results

Class	Number of correctly diagnosed images	Number of misdiagnosed images
COVID-19	924	7
NORMAL	869	29

Table 5. Model test results on modified images before using the protection methods

Class	Reference sample	Noise ($p = 0.005$)	Noise ($p = 0.05$)	Gaussian noise (1)	Gaussian noise (2)	90° rotation
0	0.000039	0.325018	0	0	0	0.000110
1	0.999999	1	0	0.101640	0	0.000000

3. PROTECTION AGAINST ML EVASION ATTACKS IN MEDICAL IMAGE ANALYSIS

3.1. Pilot Study

The Python 3.7 interpreter and the TensorFlow library [23] were used for the experiment. A simple convolutional neural network was trained, and CT images of people diagnosed with COVID-19 were taken as training and test data [24–27]. The number of training and test sets are presented in Table 3.

Table 4 presents the results of neural network training.

The accuracy of the trained model was 98.03%.

Further, noise of varying intensity was added to each of the samples, one of which is attributed to the COVID class, and the other to the NORMAL one. The first one looks like small inclusions, and the second one is Gaussian noise, more homogeneous.

Two images were generated for each type of noise: less intense and more intense. To overlay more intense Gaussian noise, several iterations of adding noise were made. In the tables below, less intense noise is indicated under number 1 in brackets; more intense noise is under number 2 in brackets. The original image was also rotated 90°.

As a result, the images obtained by changing the sample of the COVID class were also attributed to this class (the probability of classifying the sample as NORMAL only slightly increased, but very slightly). Images resulting from changing the NORMAL class sample were attributed to the COVID class. Here the probability of classification was significantly reduced.

The results are shown in Table 5, which presents the probabilities of classifying, where 0 is COVID and 1 is NORMAL. Let us note that, in this case, the noise had almost no effect on the sample that belongs to class 0, but it greatly changed the result for the sample that originally belonged to class 1. Such results may be due to the specifics of the image data: if you look at the sample, you will notice that class 0 is characterized by large bright areas and blurring (and of different intensity), which is not typical for the samples of class 1.

After that, the protection methods discussed earlier were applied to this model and data set.

3.2. Analysis of Results

The results are presented in Table 6. The cells show the probability that the ML system will attribute the image to each of the classes (COVID and NORMAL). Incorrect classification results are highlighted in red, and marginal results are in blue. Table 7 provides a comparative analysis of these protective methods based on this experiment.

Table 6. Results from implementing the protection methods

Method	Model accuracy	Class	Source sample	Salt and pepper noise (1a)	Salt and pepper noise (1b)	Gaussian noise (2a)	Gaussian noise (2b)	90 ° rotation (3)
Adversarial learning	94.59%	COVID (0)	0.0	0.0	0.0	0.0	0.0	0.0
		NORMAL (1)	1.0	1.0	1.0	1.0	1.0	0.0
Randomization (changing data)	75.23%	COVID (0)	0.0	0.0	0.0	0.0	0.0	0.0
		NORMAL (1)	0.932	0.0	0.0	0.0	0.0	0.640
Randomization (noise pollution)	85.24%	COVID (0)	0.056	0.117	0.047	0.236	0.250	0.273
		NORMAL (1)	0.878	0.751	0.530	0.928	0.934	0.082
Noise reduction (autoencoder)	98.03%	COVID (0)	0.015	0.015	0.017	0.015	0.018	0.045
		NORMAL (1)	0.674	0.659	0.605	0.645	0.564	0.034
Noise reduction (data compression 224 × 224)	98.03%	COVID (0)	0.0	0.0	0.0	0.0	0.0	0.0
		NORMAL (1)	0.999	0.999	0.999	0.999	0.999	0.0
Noise reduction (data compression 128 × 128)	98.03%	COVID (0)	0.0	0.0	0.0	0.0	0.0	0.0
		NORMAL (1)	0.999	0.999	0.994	0.999	0.998	0.0

Table 7. Comparative analysis of the protection methods

Protection method	Change in the original model architecture	Loss in the model accuracy	Increasing model training time	Increasing training sample size	Sample protection 1a	Sample protection 1b	Sample protection 2a	Sample protection 2b	Sample protection 3
Adversarial learning	No	+–	+–	+–	Yes	Yes	Yes	Yes	No
Random data change	No	+++	+++	+++	No	No	No	No	Yes
Noise pollution	Yes	++–	+–	—	Yes	Yes	Yes	Yes	No
Noise reduction (autoencoder)	No	No	—	—	Yes	Yes	Yes	Yes	No
Noise reduction (data compression)	No	No	—	—	Yes	Yes	Yes	Yes	No

—No changes, +– minor changes, ++— adequate changes, +++ significant changes.

As is shown in the experimental study, all considered protection methods can effectively cope with the evasion attack; however, each of them has its own drawbacks. Typically, they are associated with a decrease in the model accuracy or the inability to cope with adversarial samples in all cases.

For all methods that deal effectively with noisy samples (adversarial learning, noise reduction, autoencoder noise suppression, and compression), the classification of an inverted image is still a problem. Therefore, these methods should be combined with the randomization method, which can be used to make the training sample more representative.

As is shown by practical results, the noise pollution method is very sensitive to the choice of coefficients, which complicates the implementation of effective protection. The active method using an autoencoder requires more computing power, but is quite efficient. The active method by means of input data compression is also efficient, and vice versa, it does not require much computing power.

Thus, several protection methods belonging to both proactive and active methods should be combined. These should include increasing the representativeness of the training sample, adding some noise to the model layers during the training phase, and compressing the test data or using an autoencoder (which will be the second protection stage). The disadvantage of using multiple protection methods is, of course, increased system runtime, increased complexity of system design and operation, and reduced model accuracy, depending on the quality of sampling in the development of protective components.

CONCLUSION

In this work, AML attacks were investigated and their systematization was carried out. It was shown that the most dangerous, from the point of view of the authors, are evasion attacks. As a result of such effects, an attacker can influence medical procedures, and therefore, the health of people. These attacks allow an attacker to get a false positive from the system by introducing small perturbations into the test sample. For this type of attack, the main protection methods were also analyzed and their features, advantages, and disadvantages were identified, which showed the lack of a universal approach to solving the task.

One distinctive feature of medical images is that they are in most cases black-and-white, which makes it possible to introduce noise pollution that is imperceptible to the human eye. Often, the pathological areas are more "exposed" (which can be seen when examining images of healthy and sick people), which gives an advantage for the attacker: in this way, the desired sample can be attributed to another class.

As part of the study, several protective methods from the class of active and proactive protection methods were considered, namely, adversarial learning, randomization, and noise reduction. The advantages and disadvantages were highlighted in each of the methods. The experimental results showed that the defensive method either greatly affects the accuracy of the final model or is not able to cope with all types of attacks, which requires their combination.

The considered protection methods can effectively cope with an evasion attack with an accurate selection of parameters, as well as with a combination of several protection methods. The shortcomings of these approaches were identified: an increase in the training sample and requirements for computing power, a slight decrease in the accuracy of the original model, and a requirement to change the model architecture to implement a security solution.

ACKNOWLEDGMENTS

Project results are achieved using the resources of supercomputer center of Peter the Great St.Petersburg Polytechnic University—SCC Polytechnichesky (www.spbstu.ru).

FUNDING

The research is funded by the Ministry of Science and Higher Education of the Russian Federation under the strategic academic leadership program "Priority 2030" (agreement 075-15-2021-1333 dated November 30, 2021).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Ma, X., Niu, Yu., Gu, L., Wang, Yi., Zhao, Yi., Bailey, J., and Lu, F., Understanding adversarial attacks on deep learning based medical image analysis systems, *Pattern Recognit.*, 2020, vol. 110, p. 107332. https://doi.org/10.1016/j.patcog.2020.107332
2. Hospital viruses: Fake cancerous nodes in CT scans, created by malware, trick radiologists, *The Washington Post*, 2019. https://www.washingtonpost.com/technology/2019/04/03/hospital-viruses-fake-cancerous-nodes-ct-scans-created-by-malware-trick-radiologists/. Cited February 15, 2021.
3. Pitropakis, N., Panaousis, E., Giannetsos, T., Anastasiadis, E., and Loukas, G., A taxonomy and survey of attacks against machine learning, *Comput. Sci. Rev.*, 2019, vol. 34, p. 100199. https://doi.org/10.1016/j.cosrev.2019.100199

4. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D., Adversarial attacks and defences: a survey, 2018. arXiv:1810.00069 [cs.LG]
5. Barreno, M., Nelson, B., Sears, R., Joseph, A.D., and Tygar, J.D., Can machine learning be secure?, *ASIACCS '06: Proc. 2006 ACM Symp. on Information, Computer and Communication Security*, Taipei, Taiwan, 2006, New York: Association for Computing Machinery, 2006, pp. 16–25. https://doi.org/10.1145/1128817.1128824
6. Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., and Leung, V.C.M., A survey on security threats and defensive techniques of machine learning: A data driven view, *IEEE Access*, 2018, vol. 6, pp. 12103–12117. https://doi.org/10.1109/ACCESS.2018.2805680
7. Finlayso, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., and Kohane, I.S., Adversarial attacks on medical machine learning, *Science*, 2019, vol. 363, no. 6433, pp. 1287–1289. https://doi.org/10.1126/science.aaw4399
8. Taghanaki, S.A., Das, A., Hamarneh, G., Vulnerability analysis of chest x-ray image classification against adversarial attacks, *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, Stoyanov, D., Taylor, Z., Kia, S.M., Eds., Lecture Notes in Computer Science, vol. 11038, Cham: Springer, 2018, pp. 87–94. https://doi.org/10.1007/978-3-030-02628-8_10
9. Voynov, D.M. and Kovalev, V.A., Experimental assessment of adversarial attacks to the deep neural networks in medical image recognition, *Informatika*, 2019, vol. 16, no. 3, pp. 14–22.
10. Hirano, H., Minagi, A., and Takemoto, K., Universal adversarial attacks on deep neural networks for medical image classification, *BMC Med. Imaging*, 2021, vol. 21, p. 9. https://doi.org/10.1186/s12880-020-00530-y
11. Ren, K., Zheng, T., Qin, Z., and Liu, X., Adversarial attacks and defenses in deep learning, *Engineering*, 2020, vol. 6, no. 3, pp. 346–360. https://doi.org/10.1016/j.eng.2019.12.012
12. Tramer, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P., Ensemble adversarial training: attacks and defenses, *6th Int. Conf. on Learning Representations, ICLR 2018–Conf. Track Proc.*, Vancouver, 2018.
13. Liu, X. and Hsieh, Cho-J., Rob-GAN: generator, discriminator, and adversarial attacker, *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, 2019, IEEE, 2019, pp. 11226–11235. https://doi.org/10.1109/CVPR.2019.01149
14. Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A., Mitigating adversarial effects through randomization, *6th Int. Conf. on Learning Representations, ICLR 2018–Conf. Track Proc.*, Vancouver, 2018.
15. *Liu, X., Cheng, M., Zhang, H., and Hsieh, Cho-J.,* Towards robust neural networks via random self-ensemble, *Computer Vision–ECCV 2018*, Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., Eds., Lecture Notes in Computer Science, vol. 11211, Cham: Springer, 2018, pp. 381–397. https://doi.org/10.1007/978-3-030-01234-2_23
16. Dhillon GS, Azizzadenesheli K, Lipton ZC, Bernstein J, Kossaifi J, Khanna A, and Anandkumar, A., Stochastic activation pruning for robust adversarial defense, *6th Int. Conf. on Learning Representations, ICLR 2018–Conf. Track Proc.*, Vancouver, 2018.
17. Xu, W., Evans, D., and Qi, Y., Feature squeezing: detecting adversarial examples in deep neural networks, *Network and Distributed Systems Security Symp. (NDSS)*, San Diego, Calif., 2018. https://doi.org/10.14722/ndss.2018.23198
18. Samangouei, P., Kabkab, M., and Chellappa, R., Defense-GAN: protecting classifiers against adversarial attacks using generative models, *6th Int. Conf. on Learning Representations, ICLR 2018–Conf. Track Proc.*, Vancouver, 2018.
19. Shen, S., Jin, G., Gao, K., and Zhang, Y., APE-GAN: Adversarial perturbation elimination with GA, 2017. arXiv:1707.05474 [cs.CV]
20. Carlini, N. and Wagner, D., Towards evaluating the robustness of neural networks, *Proc. 2017 IEEE Symp. on Security and Privacy (SP)*, San Jose, Calif., 2017, IEEE, 2017, pp. 39–57. https://doi.org/10.1109/SP.2017.49
21. Meng, D. and Chen, H., MagNet: A two-pronged defense against adversarial examples, *CCS '17: Proc. 2017 ACM SIGSAC Conf. on Computer and Communications Security*, Dallas, 2017, New York: Association for Computing Machinery, 2017, pp. 135–147. https://doi.org/10.1145/3133956.3134057
22. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., and Zhu, J., Defense against adversarial attacks using high-level representation guided denoiser, *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, 2018, IEEE, 2018, pp. 1778–1787. https://doi.org/10.1109/CVPR.2018.00191
23. TensorFlow library. https://www.tensorflow.org/. Cited January 25, 2021.

24. Curated chest X-ray image dataset for COVID-19 detection, *Kaggle*. https://www.kaggle.com/unaissait/curated-chest-xray-image-dataset-for-covid19?select=Curated+X-Ray+Dataset. Cited February 20, 2021.
25. Chest Xray for COVID-19 detection, *Kaggle*, https://www.kaggle.com/fusicfenta/chest-xray-for-covid19-detection/. Cited February 20, 2021.
26. COVID-19 chest X-ray image dataset, *Kaggle*. https://www.kaggle.com/alifrahman/covid19-chest-xray-image-dataset/. Cited February 20, 2021.
27. COVID-19 radiography database, *Kaggle*. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database/. Cited February 20, 2021.
28. Zegzhda, D.P., Pavlenko, E., and Shtyrkina, A., Cybersecurity and control sustainability in digital economy and advanced production, *The Economics of Digital Transformation*, Devezas, T., Leitão, J., Sarygulov, A., Eds., Studies on Entrepreneurship, Structural Change and Industrial Dynamics, Cham: Springer, 2021, pp. 173–185. https://doi.org/10.1007/978-3-030-59959-1_11
29. Dakhnovich, A., Moskvin, D., and Zegzhda, D., An approach for providing industrial control system sustainability in the age of digital transformation, *IOP Conf. Ser.: Mater. Sci. Eng.*, 2019, vol. 497, p. 012006. https://doi.org/10.1088/1757-899X/497/1/012006
30. Fatin, A.D., Pavlenko, E.Yu., and Poltavtseva, M.A., A survey of mathematical methods for security analysis of cyberphysical systems, *Autom. Control Comput. Sci.*, 2020, vol. 54, no. 8, pp. 983–987. https://doi.org/10.3103/S014641162008012X

Translated by A. Kolemesin