Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Adversarial attacks in radiology – A systematic review

Vera Sorin [a,b,*], Shelly Soffer [c,d], Benjamin S. Glicksberg [e,f], Yiftach Barash [a,b], Eli Konen [a,b], Eyal Klang [a,b,g]

[a] *Department of Diagnostic Imaging, Sheba Medical Center, Ramat-Gan, Israel*
[b] *Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel*
[c] *Internal Medicine B, Assuta Medical Center, Ashdod, Israel*
[d] *Ben-Gurion University of the Negev, Be'er Sheva, Israel*
[e] *Hasso Plattner Institute for Digital Health at Mount Sinai, Department of Genetics and Genomic Sciences, New York, NY, USA*
[f] *Icahn School of Medicine at Mount Sinai, New York, NY, USA*
[g] *Sami Sagol AI Hub, ARC, Sheba Medical Center, Ramat-Gan, Israel*

ARTICLE INFO

Keywords:
Adversarial attacks
Deep learning
Radiology
Cybersecurity

ABSTRACT

Purpose: The growing application of deep learning in radiology has raised concerns about cybersecurity, particularly in relation to adversarial attacks. This study aims to systematically review the literature on adversarial attacks in radiology.

Methods: We searched for studies on adversarial attacks in radiology published up to April 2023, using MEDLINE and Google Scholar databases.

Results: A total of 22 studies published between March 2018 and April 2023 were included, primarily focused on image classification algorithms. Fourteen studies evaluated white-box attacks, three assessed black-box attacks and five investigated both. Eleven of the 22 studies targeted chest X-ray classification algorithms, while others involved chest CT (6/22), brain MRI (4/22), mammography (2/22), abdominal CT (1/22), hepatic US (1/22), and thyroid US (1/22). Some attacks proved highly effective, reducing the AUC of algorithm performance to 0 and achieving success rates up to 100 %.

Conclusions: Adversarial attacks are a growing concern. Although currently the threats are more theoretical than practical, they still represent a potential risk. It is important to be alert to such attacks, reinforce cybersecurity measures, and influence the formulation of ethical and legal guidelines. This will ensure the safe use of deep learning technology in medicine.

1. Introduction

Deep learning technology is revolutionizing healthcare, particularly within the radiology field [1–2]. Numerous applications of these models are being rapidly developed and implemented, with some already receiving FDA approval [3–5]. The technology is increasingly being integrated into practice [6–9]. However, with this technological advancement comes a concurrent rise in potential cybersecurity threats. One type of cyber-attack termed adversarial attack is gaining interest [10–11]. These attacks involve subtle manipulations of existing data or the creation of deceiving "fake" data, aimed at causing algorithmic errors. These manipulations are often imperceptible to humans, but they can mislead deep learning models, leading to inaccurate outputs. Notably, the manipulation involved in adversarial attack requires meticulous calculations designed to exploit the specific algorithm's vulnerabilities, a process which could require full access to the algorithm, or extensive trial-and-error attempts.

In the context of radiology, adversarial attacks could be designed to alter medical imaging data in a way that causes deep learning algorithms to misinterpret the images. This could potentially lead to inaccurate diagnoses based on manipulated data. This aim of this study is to review the literature on adversarial attacks in radiology, and provide explanation to some basic concepts related to the technology.

2. Methods

We searched the literature on adversarial attacks in radiology using the databases MEDLINE and Google Scholar. Studies published up to

[*] Corresponding author at: Department of Diagnostic Imaging, Chaim Sheba Medical Center, Emek Haela St. 1, Ramat Gan 52621, Israel.
E-mail address: verasrn@gmail.com (V. Sorin).

https://doi.org/10.1016/j.ejrad.2023.111085
Received 12 May 2023; Received in revised form 4 August 2023; Accepted 4 September 2023
Available online 9 September 2023

April 2023 were included. The specific keywords and search terms we employed were: ‘adversarial’, ‘attack’, ‘perturbation’, ‘example’, ‘medicine’, ‘medical imaging’, ‘radiology’, ‘reports’, ‘medical records’, ‘health records’, ‘EHR’, and ‘medical notes’. These search terms were combined, with the term ‘adversarial’ included in all searches and paired with ‘attack’, ‘perturbation’, or ‘example’. The terms associated with medical reports such as ‘radiology’, ‘report’, ‘medical records’ and so on, were also paired using appropriate Boolean operators (AND, OR). The initial search yielded 108 studies.

In addition to database searching, we manually searched the reference lists of relevant studies, including key studies in this topic that were not specifically related to radiology, for any additional studies that may have been missed during the initial search. This process resulted in the retrieval of additional 14 studies.

The inclusion criteria for our study were English language publications that describe adversarial attacks on radiology-related applications, specifically those targeting deep learning algorithms. Following the initial title and abstract screening, full-texts of potentially relevant papers were retrieved for a more detailed evaluation. We excluded papers that described data manipulation with the purpose of fooling humans, or algorithms unrelated to deep learning. This exclusion was extended to papers focusing on generative adversarial networks (GANs) for data manipulation to fool humans [3,12], as these are not explicitly adversarial attacks and were thus considered outside the scope of our current review.

Two reviewers (VS, EK) independently screened the titles and abstract of the articles resulting from the search, and any differences in their assessments were resolved through discussion to reach a consensus. Subsequently, the reviewers screened selected articles’ full-text for final inclusion. Ultimately, a total of 22 publications were included in this review. Fig. 1 presents a flow diagram of the screening and inclusion process.

The algorithms’ performance characteristics are evaluated based on the area under the receiver operating characteristic curve (AUC), accuracy, and F1 score. Some studies reported attack success rate (percentage of incorrect results). We used the individual performance metrics reported by each study.

3. Results

3.1. Adversarial attacks

Adversarial attacks deceive models by subtly manipulating input data or generating convincing fake data [10]. These alterations can be imperceptible to the human eye. While adversarial attacks primarily target algorithms, some are also designed to deceive physicians, making them a significant concern in healthcare. Fig. 2 provides a simplified illustration of adversarial perturbation, where minor modifications to an original medical image intend to fool a deep learning algorithm. By inducing misclassification of data, adversarial attacks can potentially lead to incorrect medical assessments, diagnoses, or treatment selections.

Adversarial attacks are classified into targeted and untargeted attacks. *Targeted attacks* aim to misclassify data into a specific, preselected class. In radiology this can create high-risk scenarios, such as misdiagnosing a benign pulmonary nodule as malignant. *Untargeted attacks* cause misclassification of data into any class other than the true one, potentially resulting in a spectrum of errors (e.g., classifying a pulmonary nodule as probably benign, suspicious, or very suspicious). Hirano et al. [13] employed both types of attacks on a chest X-ray classification algorithm, highlighting the high misclassification rate of normal chest X-rays as pneumonia through targeted attacks [13].

Attacks are also classified into white-box and black-box, based on the attacker’s knowledge of the algorithm. In *white-box* attacks, the attacker has full access to the algorithm’s parameters, while *black-box* attacks involve limited or no knowledge of these parameters. The perturbations

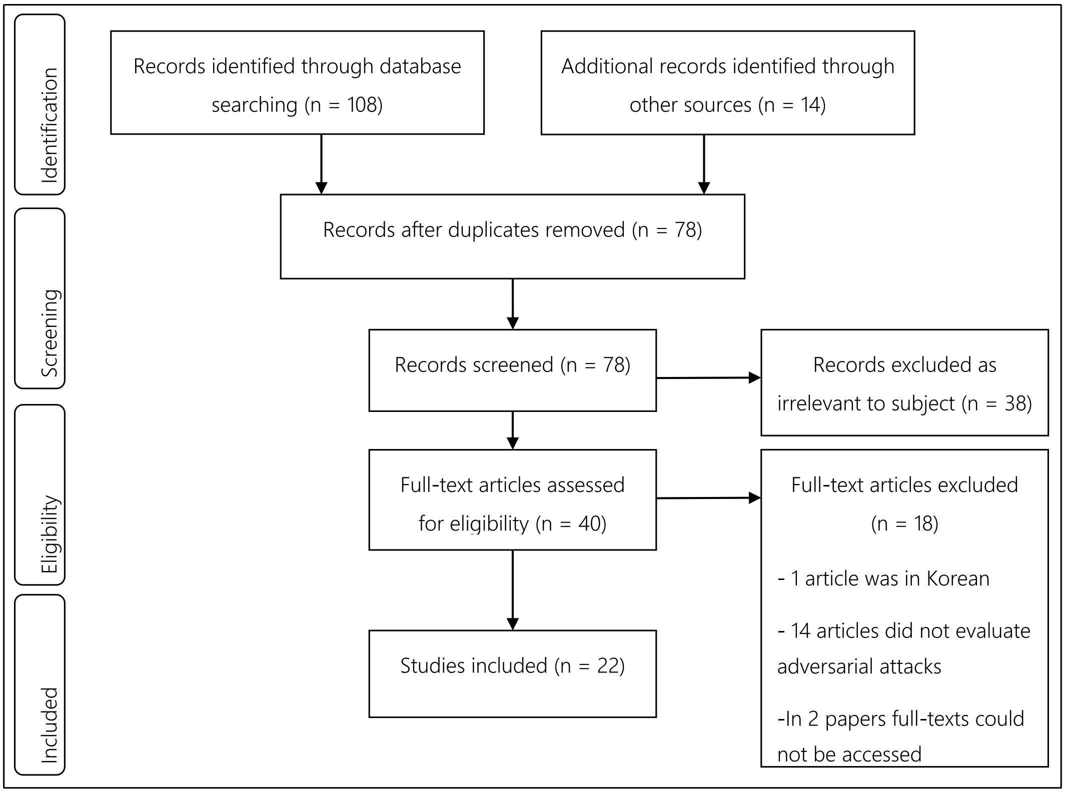
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 108)
Additional records identified through other sources (n = 14)
Records after duplicates removed (n = 78)
Records screened (n = 78)
Records excluded as irrelevant to subject (n = 38)
Full-text articles assessed for eligibility (n = 40)
Full-text articles excluded (n = 18)
- 1 article was in Korean
- 14 articles did not evaluate adversarial attacks
-In 2 papers full-texts could not be accessed
Studies included (n = 22)

Fig. 1. Flow Diagram of the Inclusion Process. Flow diagram of the search and inclusion process based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

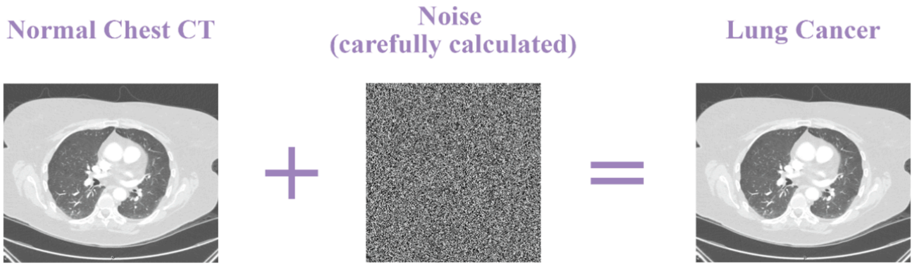

Fig. 2. **A Simplified Example of Adversarial Attack Generation.** In this example, an imperceptible carefully calculated noise is added to a normal chest CT examination. As a result, the software will categorize the apparently normal examination as one with a (nonexistent) pulmonary mass highly suspicious for lung cancer. This way, causing misclassification with potentially significant consequences to the patient.

in adversarial attacks are typically calculated using the gradient of the loss function with respect to the input, which essentially indicates how to change the input to increase the loss and cause misclassification. This process requires knowing the model's parameters and how to compute this gradient, which is characteristic to *white-box* attacks where the attacker has full access to the model. Conversely, *Black-box* attacks involve querying the model with different inputs and using the model's outputs to estimate this gradient, hence these methods may require more attempts to be successful. The threat potential from both types was seen in our review, with fourteen studies focusing on white-box attacks, three on black-box attacks, and five implemented both types. A study by Taghanaki et al. [14] applied both types of attacks on a chest X-ray classification system, highlighting the need for robust defense mechanisms.

3.2. *Adversarial attacks in radiology*

This review includes 22 studies that examined adversarial attacks in radiology, with their results summarized in Table 1. Eleven of these studies assessed adversarial attacks on deep learning models for chest X-ray classification [13–23]. Notably, Finlayson et al. [21,24] generated attacks that effectively induced misdiagnoses of pneumothorax, producing incorrect outcomes in almost every case, sometimes with 0 % accuracy and 100 % confidence [21,24]. Similarly, Hirano et al. [18] directed attacks against a COVID-19 classification model for chest X-rays, resulting in misclassification for over 90 % of the cases [18].

Ma et al. [20] suggested that medical networks might be more vulnerable to attacks than natural image networks (ImageNet), but also argued that attacks on medical networks might be easier to detect [20]. Intuitively, higher perturbation levels to an image can increase the success of an attack. Wetstein et al. [19] explored the relationship between image perturbation and the success rate of an attack, showing that while higher perturbations led to more successful attacks, they also made the manipulations more detectable, emphasizing the importance of vigilance and image quality assessment in radiology. [19].

Other targeted algorithms encompassed a range of applications, included chest CT [16,22,25–28], brain MRI [28–31], mammography [28,32], abdominal CT [33], hepatic US [34], and thyroid US [35]. For example, Paschali et al. [36] targeted algorithms for brain MRI segmentation, testing three deep learning networks with similar performance that exhibited varying resistance to attacks [36]. Byra et al. [37] conducted a black-box attack on a deep learning model for liver US image classification, using only output data to calculate the attack and resulting in misclassification for approximately half of the cases [37]. The varying resistance to attacks shown across different deep learning networks and imaging modalities indicates the complex and varied nature of the challenge posed by adversarial attacks.

While many adversarial attacks target image classification tasks, natural language processing (NLP) algorithms can also be targeted by altering texts. However, developing such attacks may be more challenging, as text modifications are generally easier for humans to detect.

4. Discussion

Adversarial attacks on deep learning algorithms in radiology have become a growing area of research. Fortunately, to the best of our knowledge, these attacks have not yet been applied to deep learning algorithms in a clinical setting. However, conventional machine learning has long been a target for such attacks, as seen in examples like spam filters being tricked into misclassifying spam emails as relevant mail through calculated alterations [11]. Adversarial attacks could reduce the accuracy of imaging classification deep learning algorithms used in radiology, a concern supported by the findings in our review.

There can be different incentives to target medical data. Attacks can influence algorithms that analyze diagnoses, medical decisions, prognosis, or financial reimbursement. Possible incentives can be economical. For example, altering medical records or billing codes. This may increase doctors' payment or secure insurance coverage for patients. Other incentives might involve undermining a competing company's algorithm or targeting individuals to affect diagnoses and medical decisions for political influence, assassination, or terrorism. Targeted cyber-attacks between nations or ethnic groups raise concerns, and ransomware attacks against healthcare systems have become increasingly common [38]. Specifically during the COVID-19 pandemic, taking advantage of hospitals' vulnerability [39–40]. Finally, attacks can erode trust in the system, a significant concern in our increasingly digital healthcare environment [41].

As AI continues to advance and become integrated into healthcare, the opportunities for adversarial attacks may increase [41–42]. Adversarial attacks could present a compelling argument for maintaining human oversight in radiology, regardless of the advancements in AI. However, even with heightened awareness, certain image modifications can go undetected by humans readers [43–44]. Although adversarial attacks in medicine remain in the research phase, the strong incentives for these attacks and the growing opportunities with the integration of AI in healthcare warrant the development of robust defenses [35,45].

Algorithms can be trained using adversarial examples (fake images simulating adversarial attacks) [46], to potentially enhance their resilience to attacks. For example, Paul et al. [25] adopted this approach in an algorithm for pulmonary nodule classification, reducing the misclassification rate and improving resistance against attacks [25]. Li et al. [29] employed an alternative strategy by incorporating contextual information into an image classification algorithm designed to predict age from brain MRI, and found that adding anatomical volumetric features increased the algorithm's resilience to attacks [29]. Other strategies include input denoising to mitigate adversarial effect, and comparing original data with the data fed into the algorithm to counteract attacks.

Defending against adversarial attacks is challenging, with mitigation efforts often involving trade-offs that can impact the algorithms' performance. When developing defenses, it is crucial to consider all possible attacks. Still, this does not mean adversarial attacks are easily applicable in real-world clinical scenarios. The perturbations designed with one

Table 1
Studies Evaluating Adversarial Attacks on Deep Learning Algorithms in Radiology.

Study (Ref)	Year	Journal	Type of Data	Type of Algorithm	Objective	Types of Attack (White / Black-Box)	Performance Before Attack	Performance Following Attack
Finlayson et al. [21,24]	2019	Science, Arxiv	Chest X-ray	CNN	Pneumothorax misclassification	Both	Accuracy 94.9 %, AUC 0.94	White box: Accuracy 0 %, AUC 0 Black box: Accuracy 15.1 %, AUC 0.014
Ma et al. [20]	2020	Pattern Recognition	Chest X-ray	CNN	Misclassification of chest X-rays (no-finding/ pneumothorax/ mass/nodule)	White-box	Accuracy 94.0 %, AUC 0.61	Accuracy 0 %, AUC 0
Paul et al. [25]	2020	2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)	Chest CT	CNN	Misclassify lung nodules	Both	Accuracy 84.4–87.3 %; AUC 0.89–0.9	White box: Accuracy 62.4–66.7 %; AUC 0.41–0.53*Black box: Accuracy 76.3–77.2 %; AUC 0.77–0.82
Li et al. [29]	2020	Neurocomputing	Brain MRI	CNN	Erroneous prediction of individual's age from brain MRI	White-box	N/A	N/A
Li et al. [33]	2019	Advances in Computer Vision and Pattern Recognition book series	Abdominal CT	CNN	Incorrect pancreas segmentation	White-box	Clean model: Dice 87.8 % Adversarially-trained model: Dice 79.09 %	Clean model: Dice 2.0 % Adversarially-trained model: Dice 65.98 %
Taghanaki et al. [14]	2018	Lecture Notes in Computer Science, vol 11038. Springer, Cham.	Chest X-ray	CNN	Misclassification of chest X-rays (pneumothorax/ mass/ cardiomegaly/ effusion/ pneumonia)	Both	Accuracy 73 %; AUC 0.77	White box: accuracy 0 %, AUC 0; Black box: accuracy 51 %, AUC 0.49
Wetstein et al. [19]	2020	Arxiv	Chest X-ray	CNN	Chest X-ray misclassification (14 pathological labels, not detailed)	White-box	AUC 0.80	AUC 0.54
Paschali et al. [30]	2018	Arxiv	Brain MRI	CNN	Incorrect brain segmentation maps	Black-box	Accuracy 71–81 % Dice 0.84–0.86	Accuracy 56–64 % Dice 0.45–0.67
	Mirsky et al. [43]	2019	USENIX Security Symposium	Chest CT	CNN	Lung cancer misclassification	White-box	The attack algorithm added and removed pulmonary nodules from images. The success of the attack when testing a CNN algorithm was 100 %. When testing radiologists, attack success ranged from 20–100 %.
Hirano et al. [13]	2021	BMC Med Imaging.	Chest X-ray	CNN	Pneumonia misclassification	White-box	Accuracy 98.4 %	Non-targeted attacks success rate 50–81.7 %; Targeted attacks success rate 92–98.3 %
Byra et al. [34]	2020	Arxiv	Hepatic ultrasound images	CNN	Fatty liver misclassification	Black-box	AUC 0.84, accuracy 82 %	Attack success rate 48 %
Hirano et al. [18]	2020	Plos One	Chest X-ray	CNN	Misclassification of COVID-19 pneumonia	White-box	Accuracy up to 94.4 %	Non-targeted attacks success rate 85.7–87.4 %; Targeted attacks success rate 93.9–100 %
Bortsova et al. [17]	2021	Medical Image Analysis	Chest X-ray	CNN	Misclassification of chest X-rays	Black-box	AUC 0.75	AUC 0.48
Shi et al. [15]	2022	Pattern Recognition	Chest X-ray	CNN	Misclassification of chest X-rays	Both	Accuracy 91.3–93.8 %, F_1 score 0.91–0.93	White box: Accuracy 0–87.2 %, F_1 score 0–0.87 Black box: Accuracy 0–92.5 %, F_1 score 0.64–0.92
Ceyhan et al. [35]	2022	Journal of Emerging Computer Technologies	Thyroid US	CNN	Misclassification of cancer at thyroid US images	White-box	Accuracy 72–93 %	Accuracy down to 1 %

(continued on next page)

Table 1 (*continued*)

Study (Ref)	Year	Journal	Type of Data	Type of Algorithm	Objective	Types of Attack (White / Black-Box)	Performance Before Attack	Performance Following Attack
Zhou et al. [32]	2021	Nature Communications	Mammography	CNN	Misclassification of mammography images	White-box	AUC 0.82	Attack success rate 69.1 %
Pal et al. [16]	2021	Applied Sciences	Chest X-ray and CT	CNN	Misclassification of COVID-19 pneumonia	White-box	CXR: accuracy 98.1 %; CT: accuracy 84 %	CXR: accuracy decreased down to 7.41 %; CT: accuracy down to 1.33 %
Li et al. [27]	2023	Bioengineering	Chest CT	CNN	Misclassification of COVID-19 at chest CT	White-box	Accuracy 80 %	Accuracy reduced to as low as 0 %
Kotia et al. [31]	2019	Advances in Intelligent Systems and Computing book series	Brain MRI	CNN	Misclassification of brain tumors	White-box	Accuracy 81.24–87.93 %	Accuracy drop down to 12.07 %
Tripathi et al. [22]	2020	Arxiv	Chest X-ray and CT	CNN	Misclassification of COVID-19 pneumonia	White-box	CXR: accuracy up to 97.28 %; CT: accuracy up to 89.19 %	CXR: accuracy drop down to as low as 7.01 %; CT: down to as low as 7.91 %
Joel et al. [28]	2021	JCO Clinical Cancer Informatics	Chest CT, mammography, brain MRI	CNN	Misclassification of cancer suspected lesions	White-box	Accuracies 75.4 % for CT, 76.4 % for mammogram, 93.6 % for MRI	Accuracy decrease of 49.8 % for CT, 52.9 % for mammogram, 87.3 % for MRI
Kansal et al.[23]	2022	Heliyon	Chest X-ray	CNN	Misclassification of COVID-19 pneumonia	Both	Accuracy 88 %	White box: Accuracy 8 %; Black box: Accuracy 67 %

Ref – reference, CNN – convolutional neural network, AUC – area under the ROC curve, EHR – electronic health records, RNN – recurrent neural network, LSTM – long short-term memory

attack are generally specific to the model being attacked. Perturbations that cause misclassification for one model might have a different effect on a different model. Thus, attacks are not universally applicable and depend on the specifics of the targeted algorithm. Furthermore, undetected access to the PACS system, or deep learning applications, presents challenges. Practical solutions tend to focus on improving information technology (IT) security [47].

The Defense Advanced Research Project Agency (DARPA) has acknowledged the potential threat of adversarial attacks, and in response has initiated the Media Forensics (MediFor) program aimed at creating a platform to identify data manipulations [48]. Regulatory frameworks for AI medical devices have been put forth, but these measures need to place more emphasis on data security [49].

Future research should focus on further understanding the intricacies of adversarial attacks and developing defenses. In particular, there is a need to study potential attacks in black-box scenarios more extensively, given their practical relevance. Radiology departments can take several actions to enhance their security. These include regular audits of IT systems, staff training on cybersecurity risks, and ongoing evaluations of AI systems for potential vulnerabilities.

This review has several limitations. First, the inherent heterogeneity of the included studies and the variability in the measures used, precluded a comprehensive quality assessment or *meta*-analysis. Second, as adversarial attacks research is ongoing, more recent studies may have been published after our review was conducted. Finally, the studies included in this review assessed the technical feasibility of adversarial attacks with full access to the targeted algorithms, although some evaluated black-box attacks.

5. Conclusions

Adversarial attacks, particularly in radiology, are a growing concern with the increasing use of deep learning in healthcare. Although currently the threats posed by these attacks are more theoretical than practical, they still represent a potential risk. It is important to be alert to such attacks, reinforce cybersecurity measures, and influence the formulation of ethical and legal guidelines. This will ensure the safe use of deep learning technology in medicine.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

CRediT authorship contribution statement

Vera Sorin: Writing – original draft, Resources, Methodology, Investigation, Formal analysis. **Shelly Soffer:** Writing – review & editing. **Benjamin S. Glicksberg:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Yiftach Barash:** Writing – review & editing. **Eli Konen:** Writing – review & editing, Supervision, Investigation. **Eyal Klang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

No funding has been received for this work.

References

[1] S. Soffer, A. Ben-Cohen, O. Shimon, M.M. Amitai, H. Greenspan, E. Klang, Convolutional Neural Networks for Radiologic Images: A Radiologist's Guide, Radiology 290 (3) (2019) 590–606.

[2] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, et al., A guide to deep learning in healthcare, Nat. Med. 25 (1) (2019) 24–29.

[3] V. Sorin, Y. Barash, E. Konen, E. Klang, Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) – A Systematic Review, Acad. Radiol. 27 (8) (2020) 1175–1185.

[4] V. Sorin, Y. Barash, E. Konen, E. Klang, Deep Learning for Natural Language Processing in Radiology—Fundamentals and a Systematic Review, J. Am. Coll. Radiol. 17 (5) (2020) 639–648.

[5] Y. Barash, G. Guralnik, N. Tau, S. Soffer, T. Levy, O. Shimon, et al., Comparison of deep learning models for natural language processing-based classification of non-English head CT reports, Neuroradiology 62 (10) (2020) 1247–1256.

[6] E. Klang, Deep learning and medical imaging, J. Thorac. Dis. 10 (3) (2018) 1325–1328.

[7] M.D. McCradden, E.A. Stephenson, J.A. Anderson, Clinical research underlies ethical integration of healthcare artificial intelligence, Nat. Med. 26 (9) (2020) 1325–1326.

[8] X. Liu, S. Cruz Rivera, D. Moher, M.J. Calvert, A.K. Denniston, A.-W. Chan, et al., Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension, Nat. Med. 26 (9) (2020) 1364–1374.

[9] N. Bhatia, H. Trivedi, N. Safdar, M.E. Heilbrun, Artificial Intelligence in Quality Improvement: Reviewing Uses of Artificial Intelligence in Noninterpretative Processes from Clinical Decision Support to Education and Feedback, J. Am. Coll. Radiol. 17 (11) (2020) 1382–1387.

[10] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples. arXiv preprint arXiv:14126572. 2014.

[11] B. Biggio, F. Roli, Wild patterns: Ten years after the rise of adversarial machine learning, Pattern Recogn. 84 (2018) 317–331.

[12] J.M. Wolterink, A. Mukhopadhyay, T. Leiner, T.J. Vogl, A.M. Bucher, I. Išgum, Generative Adversarial Networks: A Primer for Radiologists, Radiographics 41 (3) (2021) 840–857.

[13] H. Hirano, A. Minagi, K. Takemoto, Universal adversarial attacks on deep neural networks for medical image classification, BMC Med. Imaging 21 (1) (2021).

[14] S.A. Taghanaki, A. Das, G. Hamarneh, Vulnerability analysis of chest x-ray image classification against adversarial attacks, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer, 2018, pp. 87–94.

[15] X. Shi, Y. Peng, Q. Chen, T. Keenan, A.T. Thavikulwat, S. Lee, et al., Robust convolutional neural networks against adversarial attacks on medical images, Pattern Recogn. 132 (2022), 108923.

[16] B. Pal, D. Gupta, M. Rashed-Al-Mahfuz, S.A. Alyami, M.A. Moni, Vulnerability in Deep Transfer Learning Models to Adversarial Fast Gradient Sign Attack for COVID-19 Prediction from Chest Radiography Images, Appl. Sci. 11 (9) (2021) 4233.

[17] G. Bortsova, C. González-Gonzalo, S.C. Wetstein, F. Dubost, I. Katramados, L. Hogeweg, et al., Adversarial attack vulnerability of medical image analysis systems: Unexplored factors, Med. Image Anal. 73 (2021), 102141.

[18] H. Hirano, K. Koga, K. Takemoto, Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks, PLoS One 15 (12) (2020) e0243963.

[19] S.C. Wetstein, C. González-Gonzalo, G. Bortsova, B. Liefers, F. Dubost, I. Katramados, et al. Adversarial Attack Vulnerability of Medical Image Analysis Systems: Unexplored Factors. arXiv preprint arXiv:200606356. 2020.

[20] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, et al., Understanding adversarial attacks on deep learning based medical image analysis systems, Pattern Recogn. 107332 (2020).

[21] S.G. Finlayson, J.D. Bowers, J. Ito, J.L. Zittrain, A.L. Beam, I.S. Kohane, Adversarial attacks on medical machine learning, Science 363 (6433) (2019) 1287–1289.

[22] A.M. Tripathi, A. Mishra, Fuzzy unique image transformation: Defense against adversarial attacks on deep covid-19 models. arXiv preprint arXiv:200904004. 2020.

[23] K. Kansal, P.S. Krishna, P.B. Jain, R S, P. Honnavalli, S. Eswaran, Defending against adversarial attacks on Covid-19 classifier: A denoiser-based approach. Heliyon. 2022;8(10):e11209.9.

[24] S.G. Finlayson, H.W. Chung, I.S. Kohane, A.L. Beam, Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:180405296. 2018.

[25] R. Paul, M. Schabath, R. Gillies, L. Hall, D. Goldgof, Mitigating Adversarial Attacks on Medical Image Understanding Systems. 2020:1517-21.

[26] Y. Mirsky, T. Mahler, I. Shelef, Y. Elovici, editors. CT-GAN: Malicious tampering of 3D medical imagery using deep learning. 28th {USENIX} Security Symposium ({USENIX} Security 19); 2019.

[27] Y. Li, S. Liu, The Threat of Adversarial Attack on a COVID-19 CT Image-Based Deep Learning System, Bioengineering 10 (2) (2023) 194.

[28] M.Z. Joel, S. Umrao, E. Chang, R. Choi, D.X. Yang, J.S. Duncan, et al., Using Adversarial Images to Assess the Robustness of Deep Learning Models Trained on Diagnostic Images in Oncology. JCO Clinical, Cancer Inf. 6 (2022).

[29] Y. Li, H. Zhang, C. Bermudez, Y. Chen, B.A. Landman, Y. Vorobeychik, Anatomical context protects deep learning from adversarial perturbations in medical imaging, Neurocomputing 379 (2020) 370–378.

[30] M. Paschali, S. Conjeti, F. Navarro, N. Navab, Generalizability vs. robustness: adversarial examples for medical imaging. arXiv preprint arXiv:180400504. 2018.

[31] J. Kotia, A. Kotwal, R. Bharti, Risk Susceptibility of Brain Tumor Classification to Adversarial Attacks. 2020;1061:181-7.

[32] Q. Zhou, M. Zuley, Y. Guo, L. Yang, B. Nair, A. Vargo, et al., A machine and human reader study on AI diagnosis model safety under attacks of adversarial images, Nat. Commun. 12 (1) (2021).

[33] Y. Li, Z. Zhu, Y. Zhou, Y. Xia, W. Shen, E.K. Fishman, et al. Volumetric Medical Image Segmentation: A 3D Deep Coarse-to-Fine Framework and Its Adversarial Examples. 2019:69-91.

[34] M. Byra, G. Styczynski, C. Szmigielski, P. Kalinowski, L. Michalowski, R. Paluszkiewicz, et al. Adversarial attacks on deep learning models for fatty liver disease classification by modification of ultrasound image reconstruction method. arXiv preprint arXiv:200903364. 2020.

[35] M. Ceyhan, E. Karaarslan, Measuring The Robustness of AI Models Against Adversarial Attacks: Thyroid Ultrasound Images Case Study, J. Emerg. Comput. Technol. 2 (2) (2022) 42–47.

[36] M. Paschali, S. Conjeti, F. Navarro, N. Navab, editors. Generalizability vs. robustness: investigating medical imaging networks using adversarial examples. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2018: Springer.

[37] M. Byra, G. Styczynski, C. Szmigielski, P. Kalinowski, L. Michalowski, R. Paluszkiewicz, et al., editors. Adversarial attacks on deep learning models for fatty liver disease classification by modification of ultrasound image reconstruction method, in: 2020 IEEE International Ultrasonics Symposium (IUS); 2020: IEEE.

[38] E.R. Ritenour, Hacking and Ransomware: Challenges for Institutions Both Large and Small, Am. J. Roentgenol. 214 (4) (2020) 736–737.

[39] A.M. Vivian Salama, Lauren Mascarenhas, Zachary Cohen. Several hospitals targeted in new wave of ransomware attacks, CNN, 2020.

[40] L.C. Chu, A. Anandkumar, H.C. Shin, E.K. Fishman, The Potential Dangers of Artificial Intelligence for Radiology and Radiologists, J. Am. Coll. Radiol. 17 (10) (2020) 1309–1311.

[41] O. Marques, F.C. Kitamura, Trustworthiness of Artificial Intelligence Models in Radiology and the Role of Explainability, J. Am. Coll. Radiol. (2021).

[42] Y.W. Lui, K. Geras, K.T. Block, M. Parente, J. Hood, M.P. Recht, How to Implement AI in the Clinical Enterprise: Opportunities and Lessons Learned, J. Am. Coll. Radiol. 17 (11) (2020) 1394–1397.

[43] Y. Mirsky, T. Mahler, I. Shelef, Y. Elovici, editors. CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning. USENIX Security Symposium; 2019.

[44] A.S. Becker, L. Jendele, O. Skopek, N. Berger, S. Ghafoor, M. Marcon, et al., Injecting and removing suspicious features in breast imaging with CycleGAN: A pilot study of automated adversarial attacks using neural networks on small images, Eur. J. Radiol. 120 (2019), 108649.

[45] M. Levy, G. Amit, Y. Elovici, Y. Mirsky, The security of deep learning defences for medical imaging. arXiv preprint arXiv:220108661. 2022.

[46] S. Liu, A.A.A. Setio, F.C. Ghesu, E. Gibson, S. Grbic, B. Georgescu, et al., No Surprises: Training Robust Lung Nodule Detection for Low-Dose CT Scans by Augmenting With Adversarial Attacks, IEEE Trans. Med. Imaging 40 (1) (2021) 335–345.

[47] B. Desjardins, Y. Mirsky, M.P. Ortiz, Z. Glozman, L. Tarbox, R. Horn, et al., DICOM Images Have Been Hacked! Now What? Am. J. Roentgenol. 214 (4) (2020) 727–735.

[48] M. Turek, Media Forensics (MediFor). https://www.darpa.mil/program/media-forensics.

[49] D.B. Larson, H. Harvey, D.L. Rubin, N. Irani, J.R. Tse, C.P. Langlotz, Regulatory Frameworks for Development and Evaluation of Artificial Intelligence-Based Diagnostic Imaging Algorithms: Summary and Recommendations, J. Am. Coll. Radiol. (2020).