European Journal of Radiology 120 (2019) 108649

Contents lists available at ScienceDirect

European Journal of Radiology

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

Injecting and removing suspicious features in breast imaging with CycleGAN: A pilot study of automated adversarial attacks using neural networks on small images

Anton S. Becker[a,b,*], Lukas Jendele[c], Ondrej Skopek[c], Nicole Berger[a], Soleen Ghafoor[a,d], Magda Marcon[a], Ender Konukoglu[e]

[a] Institute of Diagnostic and Interventional Radiology, University Hospital of Zurich, Zurich, Switzerland
[b] Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland
[c] Department of Computer Science, ETH Zurich, Zurich, Switzerland
[d] Department of Radiology, Memorial Sloan Kettering Cancer Center, New York City, USA
[e] Computer Vision Laboratory, Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, Switzerland

ARTICLE INFO

Keywords:
Mammography
Cancer
GAN
Cyber security

ABSTRACT

Purpose: To train a CycleGAN on downscaled versions of mammographic data to artificially inject or remove suspicious features, and to determine whether these AI-mediated attacks can be detected by radiologists.
Material and Methods: From two publicly available datasets, BCDR and INbreast, we selected 680 images with and without lesions as training data. An internal dataset (n = 302 cancers, n = 590 controls) served as test data. We ran two experiments (256 × 256 px and 512 × 408 px) and applied the trained model to the test data. Three radiologists read a set of images (modified and originals) and rated the presence of suspicious lesions on a scale from 1 to 5 and the likelihood of the image being manipulated. The readout was evaluated by multiple reader multiple case receiver operating characteristics (MRMC-ROC) analysis using the area under the curve (AUC).
Results: At the lower resolution, the overall performance was not affected by the CycleGAN modifications (AUC 0.70 vs. 0.76, p = 0.67). However, one radiologist exhibited lower detection of cancer (0.85 vs 0.63, p = 0.06). The radiologists could not discriminate between original and modified images (0.55, p = 0.45). At the higher resolution, all radiologists showed significantly lower detection rate of cancer in the modified images (0.80 vs. 0.37, p < 0.001), however, they were able to detect modified images due to better visibility of artifacts (0.94, p < 0.0001).
Conclusion: Our proof-of-concept study shows that CycleGAN can implicitly learn suspicious features and artificially inject or remove them in existing images. The applicability of the method is currently limited by the small image size and introduction of artifacts.

1. Introduction

Machine learning (ML) in medical imaging is a promising field of research, which will bring substantial changes to radiology in the coming years. Many ML studies focus on (semi) automated detection [1] or classification of cancer [2]. Most advanced ML algorithms are fundamentally opaque and as they, inevitably, find their way onto medical imaging devices and clinical workstations, we need to be aware that they may also be used to manipulate raw data and enable new ways of cyber-attacks, possibly harming patients and disrupting clinical imaging service [3].

One specific genre of ML algorithms, Generative Adversarial Networks (GANs), are of particular importance in this context. GANs are a subclass of deep learning algorithms, itself a class of algorithms within the realm of ML or artificial intelligence (AI) [4]. A GAN consists of two neural networks competing against each other: The first, a generator network (G), manipulates sample images and the second, a discriminator network (D), has to distinguish between real and manipulated samples [5]. Due to their opposed cost function, the neural networks are competing against each other in order to improve their performance (in game theory this scenario is known as a “two-person zero-sum game” [6]). Given infinite resources and time, this will theoretically result in G producing samples from the real image distribution (i.e. perfect manipulations) and D incapable of discriminating,

* Corresponding author at: Institute of Diagnostic and Interventional Radiology, UniversitätsSpital Zürich, Raemistrasse 100, CH-8091 Zürich, Switzerland.
E-mail address: anton.becker@usz.ch (A.S. Becker).

https://doi.org/10.1016/j.ejrad.2019.108649
Received 15 May 2019; Received in revised form 7 August 2019; Accepted 6 September 2019

giving each such a sample a probability of 0.5 of being either manipulated or real.

On one hand, this technique may be leveraged to isolate features of the disease and either to improve our understanding of suspicious imaging features or for teaching purposes. On the other hand, for our use-case, we hypothesized that a GAN can learn an implicit representation of what suspicious lesions in mammography look like, and alter images, so they would be misdiagnosed (normal as suspicious and vice versa) without raising a suspicion of artificial manipulation. Hence, the purpose of this study was to train a pair of GANs on mammographic data to inject or remove suspicious features and determine whether these AI-mediated attacks can be detected by radiologists.

2. Methods

All used data were either publicly available [7,8] or had received prior approval for retrospective study from the local ethics committee, who waived the need for informed consent [1].

2.1. Patient Cohorts/Datasets

From two publicly available datasets, BCDR [7] and INbreast [8], 680 mammographic images from 334 patients were selected, 318 of which exhibited potentially cancerous masses, and 362 were negative controls. We used all INbreast cases with BI-RADS 3 or greater as cancer cases, and all cases with a focal lesion and marked as “malignant” from BCDR. In the first experiment, a set of images were set aside (‘evaluation dataset’) and used in the empirical assessment and evaluation, these images were not used to train the networks (15% randomly selected images). In addition, as a test dataset for experiment two (see below), we used images from a private dataset previously published in [1] (302 cancer / 590 healthy). These images were withheld from the network during the training process and only used in the last step to generate images for the readout and test how well the network generalizes to new, unseen images. Reference standard for the in-house dataset was biopsy/histopathology for malignancies and biopsy or > 2 years of follow-up for benign findings [1]. The reference standard for BCDR and INbreast was similarly defined and is described in the respective original publications [7,8].

2.2. GAN model selection and adaptation

We view the task of injecting and removing malignant features as an image-to-image translation problem in the spirit of the recently proposed cycle-consistent GANs model (CycleGAN) [9], which aim to translate images from one distribution, e.g. normal mammograms, to another distribution, e.g. mammograms with cancer, and vice versa. We trained CycleGAN, using two pairs of generator and discriminator networks to convert suspicious breast images to normal images and back to suspicious images. Similarly, the negative controls were converted to suspicious images and then back to negative/normal images.

2.2.1. First experiment

The CycleGAN architecture was implemented in TensorFlow v1.5 [10]. A schematic of the generator network architecture is shown in Fig. 1. The discriminator network was a simple convolutional network with four layers. Images were rescaled to 256 × 256 px, normalized between −1 and +1, and augmented ten-fold by random rotation, scaling, and contrast perturbations. These are standard techniques in order to make training more memory efficient (rescaling), faster (normalization) and more robust (augmentation). The training was performed on a consumer-grade personal computer (PC) with an Nvidia (Nvidia Corp., Santa Clara, CA, USA) GeForce GTX 1070 graphics processing unit (GPU). To facilitate reproducibility and possible extension of our results by others, we provide the code, along with accompanying toy data, for the first experiment in the online repository *github.com/BreastGAN/experiment1*. It contains all the relevant hyperparameters and was designed to run out-of-the-box via Docker. We trained the network for a maximum of 160k training steps. The trained CycleGAN was applied to all the test images, after appropriately scaling them to 256 × 256 pixels.

2.2.2. Second experiment

This experiment was designed and conducted after the first readout in order to further test the limits of CycleGAN. We increased the resolution of the images to 512 × 408 p × . After an initial test run with satisfactory results, we decided to proceed without data augmentation. Due to the increased image size, we used a GPU cluster consisting of eight GeForce GTX TITAN X/Xp GPUs. We implemented CycleGAN in Tensorflow v1.12.rc2 for this experiment, to match cluster requirements, and ran the network for 70k training steps. The code and synthetic data for the second experiment can be found online: *github.com/BreastGAN/experiment2*. As the first experiment, we then applied the trained CycleGAN to all the test images after scaling their size appropriately.

2.3. Radiologist readout

2.3.1. First readout

From the first experiment, we randomly chose 30 modified and 30 original images, with 40 images in pairs (i.e. original and modified from the same patient and side/view) and 20 unpaired images (i.e. different patients). In half of the images, suspicious features had been added, in the other half they had been removed. Only images with visible masses at this resolution were considered from the original images from the respective category. Hence, very dense breasts (American College of Radiology (ACR) category D) were not selected. Breast density distribution for negative/suspicious mammographic images was 11/10 for ACR cat. A, 5/9 for cat. B and 14/11 for cat. C, respectively. Median approximate lesion diameter was 6.6 mm (range: 4.0–9.9 mm). The images were presented in random order to three radiologists (5 years of experience for the two senior readers, both general radiologists, and one PGY-6 fellow in oncologic imaging) who rated them on a 5-point Likert-like scale for the likelihood of cancer (“how likely would you recall this patient”) and also voted whether they believe the image was genuine or artificially modified. In the first readout, this was a binary indication. We did not use the BI-RADS classification due to the small image size. The radiologists were not allowed to change their ratings and were fully blinded to the purpose of the study and the distribution of suspicious vs. negative cases, i.e. they were only informed that some images had been modified “by the computer”. Reference standard for analysis of the diagnostic performance was the original label of the image, even if the image had been altered by the GAN aiming to “fool” the radiologist.

2.3.2. Second readout

In the second readout, the readers knew the results of the initial readout and that CycleGAN was used for the study. During training we observed that artifacts seemed to get more pronounced due to the increased image size, especially in the later stages of the training process (see results section below). To also test this hypothesis, we presented modified images generated after different number of training iterations (35k and 70k) and let the readers rate the artificial artifacts on a 5-point scale as well. No specific training to detect artifacts was performed.

The readers were again blinded to the distribution of samples. From step 35k (half trained) we selected 12 negative and 12 suspicious images (24 images total from the evaluation dataset, half trained) to test for differences in artifact occurrence; from step 70k (fully trained) we selected 24 negative and 24 suspicious images (48 images total from evaluation and test dataset, fully trained). Half of the images were modified and half of them were originals, and again half of them paired and the other half unpaired. Hence, the total number of images for the

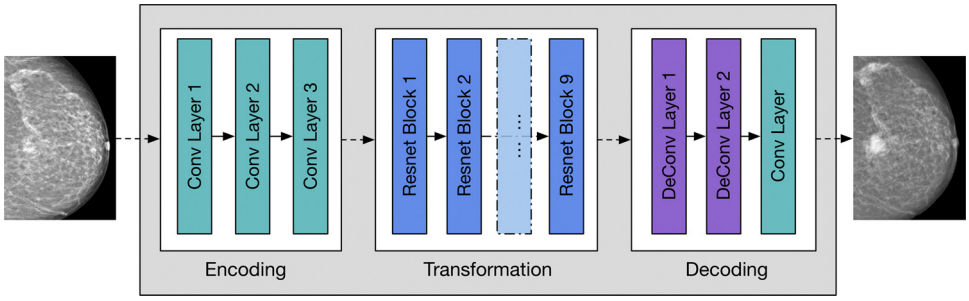

Fig. 1. Schematic diagram of the architecture of the generator network (vanilla CycleGAN implementation). The discriminator network was a simple convolutional neural network with four layers.

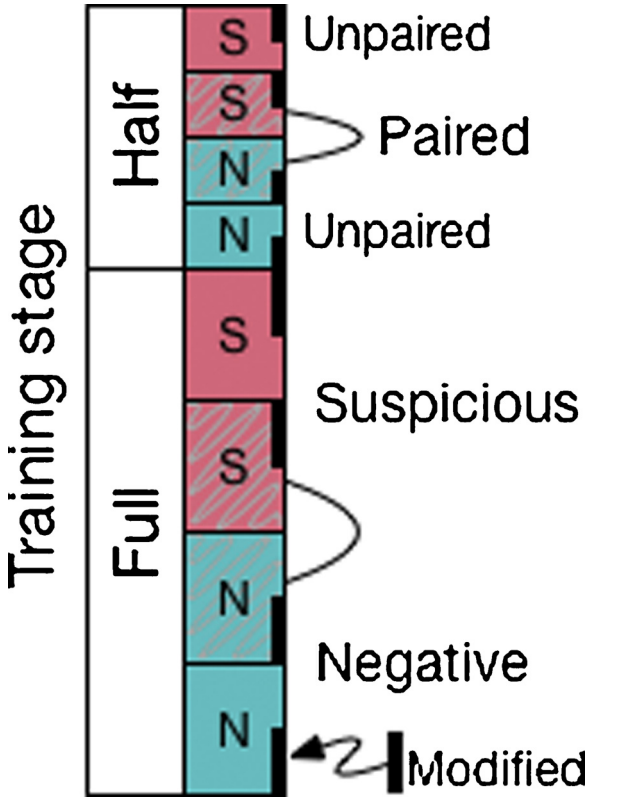

Fig. 2. Diagram illustrating the stratified image sampling for the second readout (72 images total). The stratification was similar for the first readout, without additional categorization into half and fully trained.

second readout was 72 (36 suspicious / 36 negative) as summarized in Fig. 2. Breast density distribution for negative/suspicious mammographic images was 16/15 for ACR cat. A, 12/13 for cat. B, 8/5 for cat. C and 0/3 for cat. D, respectively. Median approximate lesion diameter was 7.7 mm (range: 3.8–10.0 mm).

2.4. Statistical analysis

Statistical analysis was performed using R v.3.4.4. (R Foundation for Statistical Computing, Vienna, Austria). Continuous data were expressed as median and interquartile range (IQR). Categorical data were given in absolute counts.

Detection accuracy was assessed with receiver operating characteristic (ROC) and multiple reader multiple case (MRMC) analysis. MRMC analysis was performed with iMRMC v. 4.0.0. ROC curves for single readers were computed with the package pROC v.1.12.1. The discriminatory performance of readers was expressed as the area under the ROC curve (AUC). Averaged empirical AUC in MRMC were compared using the procedure proposed by Gallas et al. [11], single reader AUC were compared with DeLong's non-parametric test [12]. We compared lesion detection on modified vs. original images, as well as detection of CycleGAN-modifications compared to chance.

3. Results

3.1. First experiment

In a first experiment, we modified CycleGAN [9] to work with small mammographic images (256 × 256 px) from the publicly available datasets BCDR [7] and INBreast [8], running on a consumer grade PC.

Qualitatively, we noticed that at the beginning of the training, during the initial iterations, the GAN started out by first adjusting global features like contrast/brightness and then started removing or adding glandular tissue early on, thus increasing the overall breast density. Later, it would pick up skin-thickening as a suspicious feature. Finally, it would apply more focal alterations like removing or adding mass-like lesions, or morphing large, benign calcifications into fat or soft tissue masses. In general, poorly circumscribed masses would be preferentially placed on top of preexisting structures (either islets of glandular breast tissue or benign findings). We noticed that after 160k training iterations, grid-like or checkerboard-like artifacts became very prominent in the generated images, making it fairly easy for humans to spot the manipulated images. Hence, we went back to check the images for less pronounced artifacts and loaded the network with weights before iteration 160k to generate the images for the first readout.

Lesion detection in modified images: MRMC analysis revealed no significant overall difference in performance (AUC 0.70 vs. 0.76, $p = 0.67$). However, we found that in one of the experienced radiologists, the modifications introduced by CycleGAN markedly reduced diagnostic performance. The AUC of this reader dropped from 0.85 to 0.63 ($p = 0.06$) in the modified images, with regard to the original labels/classes, while the two other readers seemed unaffected, however, at a lower baseline performance (AUC 0.75 vs. 0.77 and 0.67 vs. 0.69, $p = 0.55$). Sensitivity and specificity were not significantly affected for any of the readers, ranging between 0.60-0.80 (sens.) and 0.90 (spec.) for all readers in the original images, and between 0.70-0.75 (sens.) and 0.60-0.95 (spec.), respectively.

Detection of CycleGAN-modifications: Overall, the readers could not discern between original or modified image (AUC 0.55, $p = 0.45$). However, in the per-reader analysis, the abovementioned reader could detect the CycleGAN modifications in some images (AUC = 0.66, $p = 0.008$), whereas the two other readers did not perform better than chance in this task (AUC = 0.48 and 0.50, $p = 0.59$ and 0.50).

3.2. Second experiment

Motivated by these somewhat contradictory results and a hint of success in one reader, we set out to repeat the experiment at a higher resolution. Due to the increased memory demand at the higher

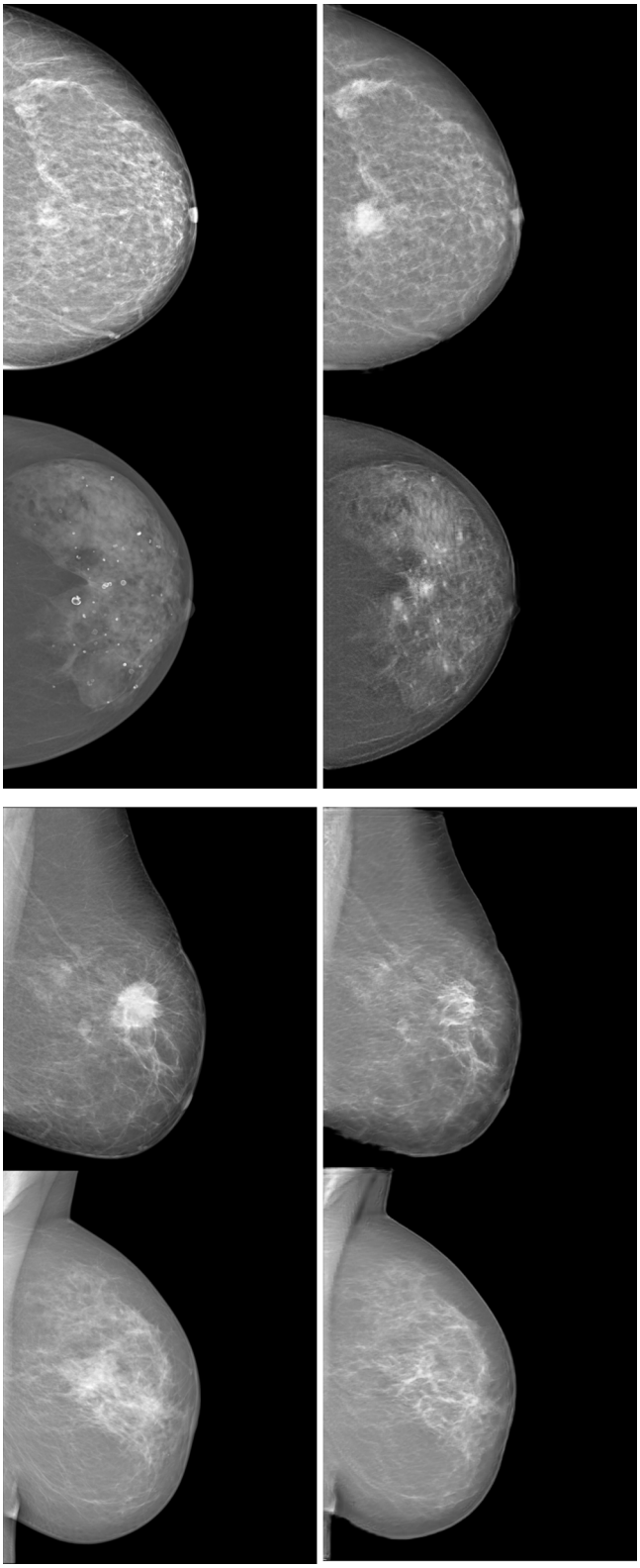

Fig. 3. Representative examples of original (left) and CycleGAN-modified images (right) from the second experiment. The top two rows are small representations of originally negative mammographic images with injected suspicious lesions, the bottom two rows are images with actual suspicious lesions, which were removed by CycleGAN. Note how in the top examples (negative to cancer), the GAN uses existing features in the image to modify in order to look suspicious (islet of breast tissue and benign macrocalcifications, respectively).

Table 1
Results of the ROC analysis comparing the detection of suspicious lesions in unmodified (original) and modified images in the second experiment. Note a significant performance drop in the modified images for all readers meaning the network succeeded to fool the reader in a substantial amount of cases. ROC = receiver operating characteristics, AUC = area under the ROC curve, Avg. = averaged AUC (MRMC). * combined p-values using the sum-z (Stouffer's) method.

	AUC originals	AUC modified	*p*-value
Avg.	0.80	0.63	*< 0.001*
Reader 1	0.78	0.69	
Reader 2	0.77	0.59	
Reader 3	0.84	0.60	
	Sensitivity orig.	Sens. mod.	*0.04**
Reader 1	0.78	0.83	0.31
Reader 2	0.89	0.72	0.10
Reader 3	0.83	0.78	0.10
	Specificity orig.	Spec. mod.	*< 0.001**
Reader 1	0.61	0.33	0.01
Reader 2	0.55	0.44	0.05
Reader 3	0.78	0.33	0.01

resolution (512 × 408 px), we ran our experiments on a dedicated GPU cluster. To test how well the network generalized to new, unseen data, we used an additional, internal test dataset from a prior study [1], which was withheld during training.

On inspection of the training monitoring, we noticed the same learning pattern as the first time, however, the gridlike artifacts were indeed more pronounced and seemed to increase already after 45–50k training iterations. The total number of images for the second readout was 72 (36 negative / 36 suspicious), a representative selection of images is shown in Fig. 3.

Lesion detection in modified images: We found that for all radiologists, the performance to discriminate between negative and suspicious (referencing the original image class) dropped significantly in modified images (AUC 0.80 vs. 0.37, $p < 0.001$; individual AUCs given in Table 1). Specificity was more affected than sensitivity ($p < 0.001$ vs. 0.04, as summarized in Table 1). A possible explanation for this is that added lesions were generally more pronounced and looked more realistic than removed ones, and artifacts after removal could still be interpreted as architectural distortions (c.f. Fig. 4).

Detection of CycleGAN-modifications: However, all radiologists could now reliably identify the modified images (AUC = 0.94, $p < 0.0001$), confirming our hypothesis that the artifacts were easier to identify at higher resolution. Identification of modifications was different in neither images from the later training stages (AUC half vs. full training = 0.93 vs. 0.94, $p = 0.83$) nor in the test dataset (evaluation vs. test = 0.93 vs. 0.95, $p = 0.68$), which did not confirm our hypothesis that the GAN would produce less artifacts at earlier training stages or in the training data (summarized in Table 2). Two representative examples of image pairs with pronounced artifacts are shown in Fig. 4.

4. Discussion

In the present study, we investigated whether a GAN can inject or remove suspicious features in a realistic way that would make modified images indistinguishable from real ones, even for radiologists, and alter the diagnosis. Our results indicate that while GANs can learn the appearance of suspicious lesions, the modification of images is currently limited by the introduction of artifacts, and the size of the images is limited by technical memory constraints.

In our initial experiment, it appeared that only one reader was influenced by the GAN modifications. After seeing a clear effect in all readers in the 2nd readout, we think that this was an effect of the higher baseline performance and not simply due to chance. The contradictory results of the first readout may also be due to different susceptibility of

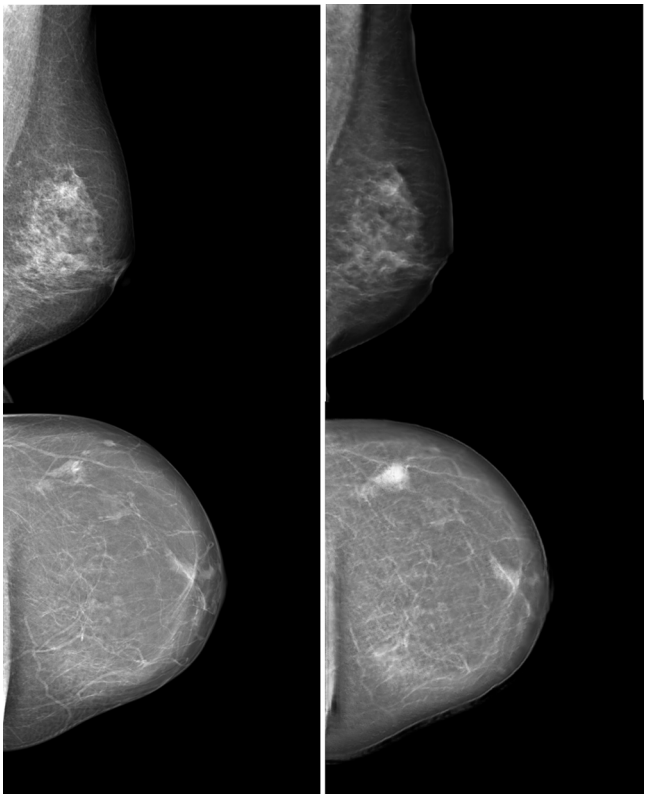

Fig. 4. Examples of images with artifacts from the second experiment/readout. Left column: Originals, right column: modified images; top row: Lesion removal, bottom row: lesion addition. Note the added artifacts in the surrounding parenchyma as well as around the edges of the skin.

Table 2
ROC analysis comparing the detection of modified images, i.e. the presence of artificial artifacts. There were no significant differences between the different subsets of data (evaluation vs. test set and half vs. fully trained), meaning CycleGAN produced artifacts even in evaluation images and at early training stages, i.e. halfway through the training process. ROC = receiver operating characteristics, AUC = area under the ROC curve, Avg. = averaged AUC (MRMC).

	AUC evaluation	AUC test	*p*-value
Avg.	0.93	0.95	0.68
Reader 1	0.93	0.91	
Reader 2	0.95	1.0	
Reader 3	0.91	0.95	
	AUC full	AUC half	*p*-value
Avg.	0.93	0.94	0.83
Reader 1	0.93	0.93	
Reader 2	1.00	0.93	
Reader 3	0.90	0.95	

the readers to 'high-frequency' signals (i.e. edges and subtle textures), which have been shown to be important in the 'gist perception' of mammographies, i.e. the first impression when viewing the image for the initial fractions of the first second [13]. The lower resolution used in the first experiment entails a loss of those signals.

In the past decades, computers have become an integral part of all clinical workflows in modern healthcare systems. This brings great advantages on one hand, i.e. better documentation, more efficient workflows [14] or new discoveries in research [15], but on the other hand, the system becomes dependent on computers and will inevitably take over some of their inherent weaknesses. The fact that such scenarios are beyond hypothetical deliberations has painfully become apparent in the recent cases where patient information in whole hospital systems was encrypted and thus rendered inaccessible by attackers who demanded a ransom payment for decryption (this particular kind of attack is called "ransomware") [16]. Moreover, these potential threats are clearly not only limited to healthcare: For example, a government investigation concluded that in the past 2016 U.S. presidential election, cyber warfare may have played a role in swaying the results in favor of a particular candidate [17].

Methods of prevention against such attacks is an important and active area of research [18]. While the need for security during storage is self-evident, our results stress the importance of security during retrieval and processing as well. To the best of our knowledge, at the moment there is no established solution for identifying artificial modifications in images made by GANs or other generative deep learning techniques. This is partly due to the fact that humans are still able to identify such generated images, even for most advanced models [19]. However, it is not far-fetched to think that these methods will soon generate images indistinguishable from reality. This article points out to this possibility for radiological images and we hope advances in prevention technologies will follow.

All modalities in a modern medical imaging department rely heavily on computers and networks, making them a prime target for cyber-attacks [3]. As machine learning or artificial intelligence (AI) algorithms will increasingly be used in the clinical routine, whether to reduce the radiation burden by reconstructing images from low-dose raw data [20,21], optimal patient positioning [22], or help diagnose diseases [1,23–25], their widespread implantation would also render them attractive targets for attacks. Exploiting vulnerabilities of deep neural networks is becoming an established field of research, yielding interesting results like the "one-pixel attack" [26], where an attacking neural network only modifies one pixel in order for the image to be misclassified. Evidently though, such an attack would not be able to fool a human observer. Hence, an important aspect of GANs is that they may in the future be able to produce realistic examples which could mislead human observers as well as machine algorithms [27]. Regarding medical imaging, we can imagine two categories of attacks: focused and generalized attacks. In a focused attack, an algorithm would be altered so it would misdiagnose a targeted person (e.g. political candidate or company executive) in order to achieve a certain goal (e.g. manipulation of election or hostile company takeover). In a generalized attack, a great number of devices would be infected with the malicious algorithm lying dormant most of the time and stochastically leading to a certain number of misdiagnoses, causing potentially fatal outcomes for the affected patients, increased cost for the whole healthcare system and — ultimately — undermining the public trust in the healthcare system. At the time of writing, however, we would argue that the technology is not yet advanced enough to make the threat of such an attack imminent. However, we think this matter deserves attention and further investigation in order to secure software/algorithms and hardware, before technology catches up.

It is worth pointing out that there are also many other possible applications of GANs apart from cyber-attacks. In a recent study, the authors investigated the use of GAN to identify features important for estimating the severity of congestive heart failure in a chest x-ray examination [28] and highlighting changes due to Alzheimer's disease [29]. Hence, GANs could be used either to discover new imaging features of a disease, for teaching purposes, or to detect biases and confounders in training datasets. Furthermore, many datasets, especially in a screening setting, are highly unbalanced, i.e. the cases of healthy individuals far outweigh the ones with cancer. GANs could be used to create more balanced datasets and thus facilitate training of other ML algorithms.

There are several limitations that also need to be mentioned. First, the introduction of grid or checkerboard artifacts as seen in our dataset is a known problem in GANs, which is related to upsampling [30]. We attribute these more perceptible artifacts at the higher resolution to two reasons: First, the higher resolution allows for finer textures and details,

and thus will require more careful modifications by the GAN in order not to distort the natural patterns occurring in the breast tissue. This is naturally more difficult and not yet perfected in current approaches, which are mostly demonstrated on much lower resolution images while clinical mammographic images have a very high resolution, about two orders of magnitude higher than the ones used in our experiments. Second, although we combined two of the largest publicly available datasets, they are still fairly small compared to datasets currently used in computer vision research, such as the ImageNet database, containing nearly 14.2 million images at the time of writing. The scarcity of data may be the limiting factor for the task at hand [12], leading to overfitting and artifacts [30]. For future research, much larger databases with mammographic images will be needed. Lastly, the different levels of experience and training background of the readers may have affected the results of the readout.

Increasing the size of the images brings another problem about: One of the most important bottlenecks for deep learning experiments in computer vision is memory. This is probably the biggest limitation of the current study. Tackling this problem is non-trivial and an active field of research. Hence, it is currently common practice in research to resize the images to a low resolution to make training feasible. We were left to choose between resizing the images, thus losing detail information, or working with small patches of full field digital mammographic images, thus losing global information. Since this was a proof-of-principle study and we were interested in whether, how and where a GAN would extract and insert features in the whole image, we chose the former trade-off. The fact that a readout with a single image patch would have been even less representative of the clinical routine was a defining point in our choice of using small, resized versions, around $1/50^{th}$ the size of the full-size mammograms (about the size of two passport size portraits). It needs to be noted that this low resolution means our findings with CycleGAN are not directly applicable to the clinical routine at this point. Compared to other images in radiology, mammographic images have the highest resolution due to the need to depict submillimeter microcalcifications. Although we could have used lower resolution images such as CT or MRI for this pilot study, we elected to use mammographic data for several reasons: First, the modality has been well established for a long time as a tool for screening and diagnostic workup, which means there are large amounts of data available from different institutions, which enabled us to compile a multi-center dataset. Second, its widespread use in screening programs makes mammography an ideal candidate for increasing cost-efficiency and performance of radiologists by ML-augmented reading [31]. As the use of other modalities for cancer screening increases, e.g. low-dose chest CT for lung cancer screening, their amenability to ML-augmented reading as well as vulnerability to cyber-attacks should be investigated as well. Although the results obtained in the current study should in principle be transferable, other network architectures or entirely different techniques will be needed for other modalities and clinical problems.

In conclusion, we could show as a proof-of-concept that a CycleGAN is capable of implicitly learning suspicious features and injecting or removing them, however, the method is limited and currently has a clear trade-off between manipulation of images and introduction of artifacts. Nevertheless, this matter deserves further study in order to shield future devices and software from AI-mediated attacks.

Declaration of Competing Interest

The authors state no relevant conflicts of interest.

References

[1] Anton S. Becker, Magda Marcon, Soleen Ghafoor, Moritz C. Wurnig, Thomas Frauenfelder, Andreas Boss, Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer, Invest. Radiol. 52 (February (7)) (2017) 434–440.

[2] Daniel Truhn, Simone Schrading, Christoph Haarburger, Hannah Schneider, Dorit Merhof, Christiane Kuhl, Radiomic versus convolutional neural networks analysis for classification of contrastenhancing lesions at multiparametric breast MRI, Radiology (November) (2018).

[3] Nilanian Dey, Amira S. Ashour, Fuqian Shi, Simon James Fong, João Manuel R.S. Tavares, Medical cyber-physical systems: a survey, J. Med. Syst. 42 (4) (2018) 74.

[4] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Deep learning, Nat. Publ. Group 521 (May (7553)) (2015) 436–444.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial nets, Advances in Neural Information Processing Systems, (2014), pp. 2672–2680.

[6] Roger B. Myerson, Game theory, Analysis of Conflict, Harvard University Press, 1997 March.

[7] Daniel C. Moura, Miguel A.Guevara Lopez, An evaluation of image descriptors combined with clinical data for breast cancer diagnosis, Int. J. Comput. Assist. Radiol. Surg. 8 (July (4)) (2013) 561–574.

[8] Ines C. Moreira, Igor Amaral, Ines Domingues, Antonio Cardoso, Maria Jo´ao Cardoso, Jaime S.˜ Cardoso, INbreast: toward a full-field digital mammographic database, Acad. Radiol. 19 (February (2)) (2012) 236–248.

[9] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, Proceedings of the IEEE International Conference on Computer Vision 2017 (2017) 2223–2232 March.

[10] Mart´ın Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, Google Brain, TensorFlow: a system for large-scale machine learning, 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI' 16) (2016) 265–284.

[11] B.D. Gallas, A. Bandos, F. Samuelson, R.F. Wagner, A framework for Random-Effects ROC analysis: biases with the bootstrap and other variance estimators, Commun. Stat. A-Theory (2009) 2586–2603.

[12] Elizabeth R. DeLong, David M. DeLong, Daniel L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics (1988) 837–845.

[13] K.K. Evans, T.M. Haygood, J. Cooper, A.M. Culpan, J.M. Wolfe, A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast, Proc. Natl. Acad. Sci. (2016) 10292–10297.

[14] Adam Yala, Regina Barzilay, Laura Salama, Molly Griffin, Grace Sollender, Aditya Bardia, Constance Lehman, Julliette M. Buckley, Suzanne B. Coopey, Fernanda Polubriaginof, Judy E. Garber, Barbara L. Smith, Michele A. Gadd, Michelle C. Specht, M. Thomas, Anthony J. Gudewicz, Guidi, Alphonse Taghian, Kevin S. Hughes, Using machine learning to parse breast pathology reports, Breast Cancer Research and Treatment, (2016), pp. 1–9 November.

[15] K.R. Chan, X. Lou, T. Karaletsos, An empirical analysis of topic modeling for mining cancer clinical notes, IEEE ICDMW 2013 (2013).

[16] Amin Kharraz, William Robertson, Davide Balzarotti, Leyla Bilge, Engin Kirda, Cutting the gordian knot: a look under the hood of ransomware attacks, Detection of Intrusions and Malware, and Vulnerability Assessment, Springer, Cham, Cham, 2015, pp. 3–24 July.

[17] Office of the Director of National Intelligence, Background to "Assessing Russian Activities and Intentions in Recent US Elections": The Analytic Process and Cyber Incident Attribution, January (2017).

[18] C.S. Kruse, B. Frederick, T. Jacobson, D.K. Monticone, Cybersecurity in healthcare: A systematic review of modern threats and trends, Technol. Health Care 25 (1) (2017) 1–10.

[19] Anton S. Becker, Michael Mueller, Elina Stoffel, Magda Marcon, Soleen Ghafoor, Andreas Boss, Classification of breast cancer from ultrasound imaging using a generic deep learning analysis software: a pilot study, Br. J. Radiol. (December) (2017) 20170576–20170578.

[20] Hu Chen, Yi Zhang, Mannudeep K. Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, Wang Ge, LowDose CT with a residual encoder-decoder convolutional neural network, IEEE Trans. Med. Imaging 36 (December (12)) (2017) 2524–2535.

[21] Moritz Schwyzer, Daniela A. Ferraro, Urs J. Muehlematter, Alessandra Curioni-Fontecedro, Martin W. Huellner, Gustav Kvon Schulthess, Philipp A. Kaufmann, Irene A. Burger, Michael Messerli, Automated detection of lung Cancer at ultralow dose PET/CT by deep neural networks initial results, Lung Cancer (November) (2018).

[22] Natalia Saltybaeva, Bernhard Schmidt, Andreas Wimmer, Thomas Flohr, Hatem Alkadhi, Precise and automatic patient positioning in computed tomography: avatar modeling of the patient surface using a 3-Dimensional camera, Invest. Radiol. 53 (November (11)) (2018) 641–646.

[23] Tero Karras, Samuli Laine, Timo Aila, A style-based generator architecture for generative adversarial networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) 4401–4410.

[24] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Y.Ng. Andrew, CheXNet: RadiologistLevel Pneumonia Detection on Chest X-Rays With Deep Learning, (2017) November arXiv.org.

[25] Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, Georgevanden Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, C.´ıan

O. Hughes, Rosalind Raine, Julian Hughes, Dawn A. Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees, Trevor Back, Peng T. Khaw, Mustafa Suleyman, Julien Cornebise, Pearse A. Keane, Olaf Ronneberger, Clinically applicable deep learning for diagnosis and referral in retinal disease, Nat. Med. 24 (September (9)) (2018) 1342–1350.

[26] Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi, One pixel attack for fooling deep neural networks, IEEE Trans. Evol. Comput. (2019).

[27] Samuel G. Finlayson, Isaac S. Kohane, Andrew L. Beam, Adversarial attacks on medical machine learning, Science 363 (March (6433)) (2019) 1287–1289.

[28] Jarrel C.Y. Seah, Jennifer S.N. Tang, Andy Kitchen, Frank Gaillard, Andrew F. Dixon, Chest radiographs in congestive heart failure: visualizing neural network learning, Radiology (November) (2018).

[29] C.F. Baumgartner, L.M. Koch, K. Can Tezcan, J. Xi Ang, E. Konukoglu, Visual feature attribution using wasserstein gans, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) 8309–8319.

[30] Augustus Odena, Vincent Dumoulin, Chris Olah, Deconvolution and checkerboard artifacts, Distill 1 (October (10)) (2016) e3.

[31] Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Jonas Teuwen, Mireille Broeders, Gisella Gennaro, Paola Clauser, et al., Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study, Eur. Radiol. (April) (2019).