Longitudinal Mammogram Exam-based Breast Cancer Diagnosis Models: Vulnerability to Adversarial Attacks

Zhengbo Zhou⋆   Degan Hao⋆   Dooman Arefan[†]   Margarita Zuley[†]
Jules Sumkin[†]   Shandong Wu⋆[†§]

⋆Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA
[†]Department of Radiology, University of Pittsburgh, Pittsburgh, PA, USA
[§]Department of Biomedical Informatics and Department of Bioengineering
University of Pittsburgh, Pittsburgh, PA, USA

Abstract. In breast cancer detection and diagnosis, the longitudinal analysis of mammogram images is crucial. Contemporary models excel in detecting temporal imaging feature changes, thus enhancing the learning process over sequential imaging exams. Yet, the resilience of these longitudinal models against adversarial attacks remains underexplored. In this study, we proposed a novel attack method that capitalizes on the feature-level relationship between two sequential mammogram exams of a longitudinal model, guided by both cross-entropy loss and distance metric learning, to achieve significant attack efficacy, as implemented using attack transferring in a black-box attacking manner. We performed experiments on a cohort of 590 breast cancer patients (each has two sequential mammogram exams) in a case-control setting. Results showed that our proposed method surpassed several state-of-the-art adversarial attacks in fooling the diagnosis models to give opposite outputs. Our method remained effective even if the model was trained with the common defending method of adversarial training.

Keywords: Breast cancer · Adversarial attack · Longitudinal model, Diagnosis.

1 Introduction

Mammography-based AI is at the forefront of medical AI research and many AI products are being translated for clinical deployment. The cyber-security of such models is therefore becoming a paramount need to ensure AI integrity. Although medical IT system is relatively closed, but cyber-attacks still occur, quite often, through infiltration of the IT systems [1] or by internal hackers [2]. For example, hospitals had to pay ransom when their data are hacked [3]. In fact, 94 % of U.S. hospitals are affected by healthcare data breaches [4], showing high vulnerability and risks to adversarial attacks. As elucidated in [5], adversarial attacks can occur with real world bad intentions for data-based ransom, insurance fraud, clinical trial effect manipulation, etc.

In clinical radiology, the established practice of comparing prior data with current data serves as a cornerstone for radiologists to perform lesion detection and diagnosis. The emergence of deep learning models has refined this process by integrating temporal sequential data, such as using multiple mammograms, magnetic resonance imaging, computed tomography, etc., as model inputs. Recent evidences indicate that these longitudinal models outperform those relying solely on a single time-point input. Cui et al. [6] shows that applying convolutional neural networks to follow-up MRI scans can effectively diagnose Alzheimer's disease. Dadsetan et al. [7] proposes a novel deep learning model on longitudinal mammogram showing superior effects compared to the model using a single exam. Lee et al. [8] leverages a Transformer decoder to incorporate prior images showing improved performance on breast cancer risk prediction.

In the biomedical domain, longitudinal models using sequential imaging scans are gaining popularity, where these models are effective by utilizing spatiotemporal relationships of longitudinal data. Adversarial attacks on models using an individual time-point scans have received considerable attention [9]. However, there has been little to none exploration into attacks on longitudinal models.

In this work, we focus on the task of diagnosing breast cancer through a Transformer decoder architecture [10] using sequential mammogram exams. We examine the susceptibility of this longitudinal model to adversarial attacks. We propose a novel attack method tailored to combine cross-entropy loss (to guide adversarial samples across the decision boundary) and distance metric learning (to modify the relationship between sequential imaging exams), aiming to fool the diagnosis models to give rise to an opposite output. We followed the attack transferring scheme and showed through experiments that, even when adversarial training is employed to enhance the model's robustness, our attacking method can still effectively degrade the model's performance. Our method also outperforms several other compared methods. Our study highlights a significant vulnerability of longitudinal models to adversarial attacks, which urges the needs of enhancing such model's safety against adversarial attacks.

The contributions of our work can be summarized as: (1) We studied a novel topic of investigating adversarial attacks on longitudinal imaging-based diagnostic models; (2) We evaluated the method of integrating cross-entropy loss with distance metric learning to implement the attack, which exhibited stronger attaching effects in leading to misclassification of breast cancer, compared to several state-of-the-art attack techniques; and (3) We showed that in distance metric learning, medical knowledge can be leveraged in selecting effective adversarial samples aiming to fool a diagnosis model.

2   Related Work

Adversarial Attack Methods: Szegedy et al [11] showed that classifiers may confidently make incorrect predictions when subjected to imperceptible perturbations. Kurakin et al. [12] showed that classifiers remain vulnerable to adversarial samples even in the physical world. Goodfellow et al. [13] propose FGSM,

a simple and efficient adversarial samples generating method. In addition, Basic Iterative Method (BIM) [14] is proposed as an extension of the Fast Gradient Sign Method (FGSM) and aims to generate stronger adversarial samples by iteratively applying FGSM with smaller perturbation steps. Benefiting from the iteration strategy, MI-FGSM is proposed by adding momentum through iterations [15]. The Projected Gradient Descent (PGD) attack [16] is another popular and powerful attack method. On the other hand, optimization-based C&W attack [17] treat the adversarial sample generation as an optimization problem.

Attack Transferability: The vulnerability of models to adversarial samples generated by other models is known as the transferability of adversarial attacks [11]. Kurakin et al. [12] conducted a study on adversarial samples within the context of ImageNet. Their found that BIM's multi-step approach is less transferable than the single-step FGSM. Zhou et al. [18] showed that enhancing the transferability of the BIM can be achieved by maximizing the distance between natural images and their adversarial counterparts within the intermediate feature maps, coupled with the addition of regularization. The transferability of a single-input model to longitudinal models is an interesting approach but not yet studied in the literature.

Defense Methods: Hinton et al. [19] introduced distillation as a method to bolster resistance against adversarial samples. The adversarial training method involves incorporating adversarial samples into the training set to enhance model robustness through retraining [12]. Tramèr et al. [20] expanded on this approach with ensemble adversarial training, augmenting training data with perturbations transferred from other models. In general, adversarial training is a common defending method for adversarial attacks.
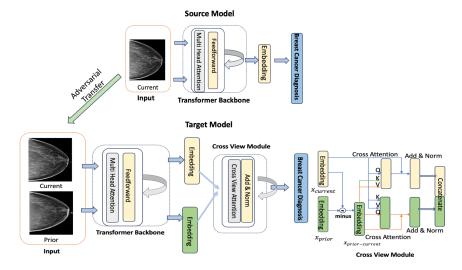

Fig. 1: Adversarial attacks transferred from a source model (a single input, Current) to a target model (two sequential input, Prior and Current) for breast cancer diagnosis.

3 Methods

3.1 Task and Model Architecture

Task. In this study, we implemented our novel method on the task of diagnosing breast cancer on mammograms (i.e., classification of cancer cases vs. normal controls). The diagnosis can be based on a single time-point exam or multiple longitudinal exams that leverage temporal information to improve diagnosis accuracy. Thus, we designed two models: Source model (using a single time-point mammogram exam, denoted as Current) and Target model (using two sequential exams, denoted as Prior and Current, respectively, which are taken at two different time points). The Source model is useful when a patient does not have Priors available. It should be pointed out that the Prior exams are normal/healthy mammograms that are stored as historical data of a patient in the electronic health records. The use of normal Priors in longitudinal models conforms to radiologists' clinical practice to improve lesion detection and diagnosis in addition to using the Current exam.

Model Architecture. As illustrated in Figure 1, in the Source model, features extracted by the backbone model are utilized for diagnosis, which also serve as the basis for generating adversarial samples to attack the target longitudinal model i.e., (the attack transferring). For the Target longitudinal model, the imaging features are derived from both the Prior and Current exams, referred to as x_{prior} and x_{current}, respectively. The inputs to the cross-view module are current exam x_{current} and the changes between the prior and current exam, i.e., $x_{\mathrm{prior\text{-}current}} = x_{\mathrm{prior}} - x_{\mathrm{current}}$. Here $x_{\mathrm{prior\text{-}current}}$ aims to capture and emphasize the changes between the two time points. Subsequently, both feature vectors are fed to the cross-view module. In this module, we use multi-head attention mechanisms to identify and stress the information in the subtracted feature ($x_{\mathrm{prior\text{-}current}}$) that is relevant to the current feature (x_{current}). The core of multi-head attention is a scaled dot-product attention[10] mechanism that computes attention scores between queries and keys, which are then used to linearly combine the associated values. In our setting, we reshape the embedded feature maps for the target phase (x_{current} or $x_{\mathrm{prior\text{-}current}}$) into a query matrix and reshape the feature maps for the source phase into a key matrix. The outcome of this attention process produces $x_{\mathrm{current\text{-}attention}}$ and $x_{\mathrm{(prior\text{-}current)\text{-}attention}}$. To form new attention-enhanced feature vectors, we integrate these representations with the original vectors, resulting in $x_{\mathrm{(prior\text{-}current)\text{-}cross}} = x_{\mathrm{prior\text{-}current}} + x_{\mathrm{(prior\text{-}current)\text{-}attention}}$ and $x_{\mathrm{current\text{-}cross}} = x_{\mathrm{current}} + x_{\mathrm{current\text{-}attention}}$. These refined feature vectors are then concatenated, serving for the ultimate prediction.

3.2 Adversarial Attack

Rationale: Our study aims to attack the Target model that makes diagnosis using the temporal relationships/changes of the Prior and Current exams. In our black-box attacking scenario, attackers do not have access to the model architecture and parameters of the Target model, particularly the mechanisms in

which how the longitudinal features are used. Attackers craft adversarial samples using the Source model, following the scheme of "attack transferring", to perform attacks to the Target model. This makes sense because attackers do not need to or may not have a longitudinal models to generate adversarial samples; instead, a Source model based on a single time-point input is much more accessible and easily available to use. Here, the attacks aim to manipulate the Current exam without considering the Prior exam, as the cancer diagnosis is primarily based on the Current exam (recall that the Prior exams are all normal/healthy and
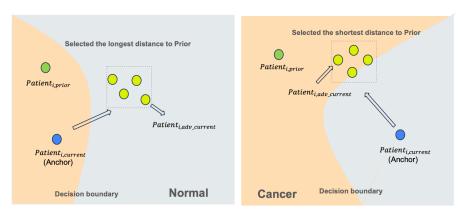

Fig. 2: Illustration of knowledge-guided selection of samples for adversarial Current: to select the one with the shortest (longest) distance to Prior for Cancer (Control) cases. Light orange represents Normal decision space, while gray represents Cancer decision space.

Adversarial Sample Generation. Let $f(s)$ denote an arbitrary deep neural network which takes s ($s \in \mathbb{R}^n$) as input and outputs the probability of classes y ($y \in \mathbb{R}^m$). We first define an adversarial sample fooling the model $f(s)$ for a chosen p-norm and noise parameter ϵ as follows:

$$\tilde{s} = \text{argmax}_{\|\tilde{s}-s\|_p \leq \epsilon} l(\tilde{s}, t) \tag{1}$$

where t and $l(\cdot, \cdot)$ denote the label of x and the loss function used to train the model respectively. In our experiments, we investigate the effects of using the cross entropy loss and distance metric learning for attacking the Target model. To optimize Equation (1), the Iterative Fast Gradient Sign Method (I-FGSM) is employed. Initially, the input x is scaled to the range into $[-1, 1]$ and we set $\tilde{x}_0 = x$. Then, we compute the gradients of loss function with respect to input x. Following this, the adversarial samples undergo several iterations of updates. In each step, we apply the sign function to the calculated gradients and then clip the updated adversarial samples to the range $[-1, 1]$ to maintain them as valid images. At every iteration, the generated samples are designed to cross the decision boundary, directed by the gradient data, and are retained for subsequent sampling. Ultimately, adversarial samples are generated by adding

the pixel differences, scaled by ϵ, between the most recently updated adversarial samples and the original input x.

Knowledge-guided Adversarial Sample Selection Method. As illustrated in Figure 2, the I-FGSM method can generate multiple adversarial samples aiming at traversing the decision boundary. We propose a novel criteria to select most impacting adversarial samples using distance metric learning. This criteria is designed based on the knowledge of: (1) the longitudinal model relies on the relationship of Prior (always normal) and Current (could be cancer or normal) exams to make a diagnosis, and (2) the Euclidean distance between Prior and Current is smaller for a normal patient (because of less variation across two normal exams), but larger for a cancer patient (because of larger variation from normal transitioning to cancer development). Specifically the criteria differ between Normal and Cancer cases: in Cancer cases, we select the adversarial sample $Patient_{i,adv\_current}$ that is closest from $Patient_{i,prior}$, while in Control cases, we choose the sample that is furthest to $Patient_{i,prior}$. For Cancer cases, the objective is to coax the model into misclassifying the adversarial sample as Normal by leveraging the proximity to a $Patient_{i,prior}$ image labeled as normal. This proximity strategy enhances the likelihood that the longitudinal model, which relies on temporal relationships for classification, erroneously assesses the sample as Normal. Conversely, for Normal cases, the intent behind selecting adversarial samples with the greatest distance from the Prior image is to challenge the model into a false diagnosis of Cancer. Given the substantial separation between the adversarial sample and a Prior image, the longitudinal model is more inclined to misclassify these as belonging to Normal cases. This novel adversarial sample selection approach takes advantages of the longitudinal model's dependencies on the relationship between $Patient_{i,prior}$ and $Patient_{i,current}$ exams.

4 Experiments

4.1 Study Cohort

This study received institutional review board approval and we used a dataset of 590 subjects in a case-control study setting, with 293 breast cancer cases and 297 breast cancer-free controls (i.e., normal/negative). Each subject has a Current mammogram exam and a Prior exam taken at approximately 1 year apart. All diagnosis outcomes are biopsy-proven and based on the Current exams. All Prior exams are normal/negative. For the cancer cases, we exclusively utilized the biopsied breast for. For the controls, either the right or left breast was randomly chosen for a subject to avoid modeling from shortcut learning. To ensure uniformity in the orientation of a breast, we applied a horizontal flip for a right-side breast to appear like a left-side breast. We specifically opted for using the craniocaudal view of the mammogram images.

4.2 Implementation Details

The dimensions of the input mammogram images were standardized to 350x400 pixels to maintain consistency across the subjects. We employed a Swin Trans-

former [21] pretrained on ImageNet as the backbone for feature extraction in both the Source and Target models. We also evaluated the effects of implementing the Source model with a VGG [22] model as the backbone architecture, whereas the Target model remains unchangedand and reported results in Supplementary materials. To prevent overfitting for both Source and Target model training, we incorporated data augmentation during the training phase, including flipping and image rotation. The model underwent training over 30 epochs, with learning rates of 5e-5 and 1e-5 identified as the optimal parameter values through experiments. For generating adversarial attacks, the iteration number was set to 15 for all iterative-based attack methods, and the perturbation size was set to 0.01 for all attacks. We performed parameter robustness analysis experiments on these two parameters (See results in Supplementary materials).

We compared the effects of our method to several methods, including C&W, FGSM, I-FGSM, MI-FGSM, and PGD. Also, we compared to two distance-guided new methods using distance loss on FGSM and I-FGSM, respectively, as described by Equation (2), where we denote x as input data, t as label, L(x) as the intermediate feature map:

$$\text{loss}_{\text{distance}}(x', t) = \begin{cases} L(x') - L(x_{\text{prior}}) & \text{if } t = 1 \\ L(x_{\text{prior}}) - L(x') & \text{if } t = 0 \end{cases} \tag{2}$$

In addition, we implemented two more novel methods that used distance metric loss to alter the relationship between Prior and adversarial Current. Specifically, we use the distance loss in Equation (2) as a penalty term aiming to more efficiently guide the search directions. We named this regularization method as 'Distance Reg.' (Equation (3)) as another way to combine the cross entropy loss and distance metric learning. We compared our proposed knowledge-guided method to this alternative method (λ is experimentally determined as 0.05).

$$\tilde{s} = \text{argmax}_{\|\tilde{s}-s\|_p \leq \epsilon} l(\tilde{s}, t) + \lambda \cdot \text{loss}_{\text{distance}}(s', t) \tag{3}$$

We utilized a patient-wise 5-fold cross-validation and Area Under the ROC Curve (AUC) reporting Mean AUC ± Standard Deviation (Std) of the various methods. We also assessed the model's performance after incorporating adversarial training as defence method, where the model underwent retraining with a combination of clean data and adversarial samples generated by BIM with perturbation size of 0.01 and batch size of 32. All computational tasks were performed on an NVIDIA TESLA V100 GPU, provided by our local supercomputing facility.

Table 1: The performance of different attacks on Source, Target without, and Target with adversarial training models. (Format: mean AUC ± std).

Attack Method	Source Model	Target Model	Target Model (Adversarial Training)
No Adversarial Attack	0.670 ± 0.017	0.704 ± 0.054	0.685 ± 0.033
FGSM [13]	0.194 ± 0.052	0.405 ± 0.045	0.584 ± 0.044
Distance-guided FGSM	0.403 ± 0.063	0.531 ± 0.040	0.634 ± 0.029
Distance Reg. FGSM	0.292 ± 0.080	0.452 ± 0.053	0.606 ± 0.031
I-FGSM [14]	0.0 ± 0.0	0.322 ± 0.056	0.573 ± 0.038
Distance-guided I-FGSM	0.039 ± 0.014	0.430 ± 0.030	0.612 ± 0.037
Distance Reg. I-FGSM	0.0 ± 0.0	0.286 ± 0.049	0.573 ± 0.031
C&W [17]	0.0 ± 0.0	0.376 ± 0.063	0.572 ± 0.035
MI-FGSM [15]	0.0 ± 0.0	0.297 ± 0.046	0.570 ± 0.043
PGD [16]	0.0 ± 0.0	0.289 ± 0.064	0.573 ± 0.031
AutoAttack [23]	0.0 ± 0.0	0.308 ± 0.042	0.585 ± 0.034
GI-FGSM [24]	0.0 ± 0.0	0.286 ± 0.04	0.587 ± 0.028
PC-I-FGSM [25]	0.0 ± 0.0	0.291 ± 0.05	0.584 ± 0.016
GRA [26]	0.0 ± 0.0	0.303 ± 0.045	0.574 ± 0.036
Our Proposed Attack	0.0 ± 0.0	**0.205** ± 0.040	**0.548** ± 0.036

Table 2: The performance of different attack methods when using VGG as the backbone for the Source model. The Target models (including the one with adversarial training) remained the same (i.e., Swin Transformer). (Format: mean AUC ± std).

Attack Method	Source Model VGG	Target Model Swin Transformer	Target Model (Adversarial Training)
No Adversarial Attack	0.66 ± 0.045	0.704 ± 0.054	0.685 ± 0.033
FGSM [13]	0.161 ± 0.055	0.546 ± 0.05	0.597 ± 0.039
Distance-guided FGSM	0.365 ± 0.043	0.545 ± 0.045	0.615 ± 0.029
Distance Reg FGSM	0.292 ± 0.08	0.452 ± 0.053	0.606 ± 0.031
I-FGSM [14]	0.0 ± 0.0	0.45 ± 0.078	0.595 ± 0.038
Distance-guided I-FGSM	0.0 ± 0.0	0.515 ± 0.079	0.603 ± 0.037
Distance Reg I-FGSM	0.0 ± 0.0	0.447 ± 0.049	0.596 ± 0.031
C&W [17]	0.0 ± 0.0	0.502 ± 0.074	0.580 ± 0.035
MI-FGSM [15]	0.0 ± 0.0	0.446 ± 0.082	0.591 ± 0.043
PGD [16]	0.0 ± 0.0	0.461 ± 0.113	0.596 ± 0.017
AutoAttack [23]	0.0 ± 0.0	0.514 ± 0.066	0.585 ± 0.019
GI-FGSM [24]	0.0 ± 0.0	0.475 ± 0.088	0.617 ± 0.011
PC-I-FGSM [25]	0.0 ± 0.0	0.53 ± 0.071	0.626 ± 0.012
GRA [26]	0.0 ± 0.0	0.464 ± 0.115	0.614 ± 0.016
Proposed Attack	0.0 ± 0.0	**0.418** ± 0.1	**0.577** ± 0.016

5 Results

As shown in Table 1, the Source and Target models show baseline AUCs of 0.670 and 0.704, respectively. The Target model has a higher AUC than the Source
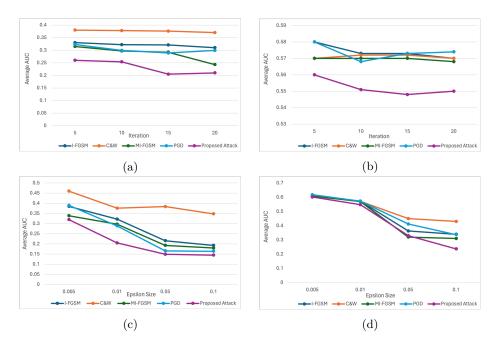
Fig. 3: **Parameter robustness analysis.** Target model performance is shown with respect to a range of values of the iteration number (A, B) and perturbation size (C, D). Adversarial training was applied in B and D.

model, indicating the usefulness of the use of Prior exams for diagnosis. With adversarial attacks, all methods lowered the AUCs to a range of 0.205 to 0.531, showing all the attacks successfully fooled the model to give opposite (AUC<0.5) or random (AUC=0.5) diagnosis. Our proposed attack led to the lowest AUC of 0.205, outperforming all the other compared methods.

Table 1 last column shows the effects of after incorporating adversarial training to retrain the Target models. It shows that the Target model achieved an AUC of 0.685 when testing with clean data, which is slightly lower than the AUC (0.704) without adversarial training – this shows a trade-off between increased adversarial robustness and slightly-decreased performance on clean data. As can be seen, the Target model becomes more resilient to all the adversarial attacks as the AUC increased to higher values (range 0.548 to 0.634) compared to those (0.205 to 0.531) without adversarial training. This partially indicates the adversarial training is useful to mitigate the attacks to certain extent, but still, the model performance remains much lower than the normal performance (0.685), which means the attacks can still substantially fool the diagnosis model. Here, again, our proposed attack method achieved the best attacking effects.

The results also show that both the cross-entropy loss (FGSM, I-FGSM), PGD and distance-guided learning (Distance-guided FGSM, Distance-guided I-FGSM) can independently degrade model performance. The Distance Regularization-

based methods perform better than the distance-guided methods, but are less effective than our proposed approach – this is potentially due to that the distance metric loss is not necessarily able to ensure that, in altering the relationship between the adversarial Current and Prior, the adversarial Current will breach the decision boundary, a task typically however can be influenced by cross-entropy loss. This suggests that our knowledge-guided sample selection method used to generate adversarial samples is a more effective method for attacks.

It should be noted that the generated adversarial samples are supposed to be able to fool the Source model (even though the real intention is to attack the Target model). This is verified in Table 1 from the low AUCs (0 means an opposite diagnosis of the entire cases) of the Source model. In addition, for the knowledge described for sample selection, our experiments also provided quantitative statistics supporting the validity of the distance knowledge: the average Euclidean distance between Prior and Current is 0.38 (smaller) and 0.52 (larger) on the normal and cancer patients, respectively.

Table 2 presents the performance of various attack methods using VGG as the backbone for the source model, while keeping the target models (including the adversarially trained model) unchanged. The results indicate that the overall attack performance patterns are consistent with those observed in Table 1, suggesting that our method is effective across different model architectures, including non-Transformer-based models.

From Fig. 3, our method consistently outperforms the other four methods (I-FGSM, C&W, MI-FGSM, PGD) across a range of parameter values, both with and without adversarial training. Notably, as the epsilon value increases—indicating stronger adversarial perturbations—AUC values for all methods decrease, as expected.

6   Conclusion

In this study, we delved into the medical imaging AI model robustness against adversarial attacks on longitudinal models, with a particular focus on breast cancer diagnosis. Our research topic is novel because studies on adversarial attacks to longitudinal models are rare, yet such models are gaining popularity in medical applications. We proposed a novel attacking method that combines cross-entropy loss and knowledge-guided distance metric learning, showing much superior effects in terms of fooling the diagnosis model, and outperforming several compared state-of-the-art methods. Our method remained effective even after incorporating the defensing method of adversarial training. Future work includes further evaluation on different deep learning structures and developing effective defense methods. Our study highlights the importance and urgency of adversarially robust medical diagnosis models towards delivering safe AI to patient care.

7 Acknowledgements

References

1. Sean Lyngaas. Cyberattack is a factor in illinois hospital's closure. *CNN*, June 2023.
2. Y Mirsky, T Mahler, I Shelef, and Y Elovici. Ct-gan: malicious tampering of 3d medical imagery using deep learning. 2019, 2019.
3. Berkeley Lovelace Jr. Hospital ceo forced to pay hackers in bitcoin now teaches others how to deal with ransomware. *CNBC*, April 2018.
4. HIPAA Journal. Healthcare data breach statistics, n.d.
5. Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.
6. Ruoxuan Cui, Manhua Liu, Alzheimer's Disease Neuroimaging Initiative, et al. Rnn-based longitudinal analysis for diagnosis of alzheimer's disease. *Computerized Medical Imaging and Graphics*, 73:1–10, 2019.
7. Saba Dadsetan, Dooman Arefan, Wendie A Berg, Margarita L Zuley, Jules H Sumkin, and Shandong Wu. Deep learning of longitudinal mammogram examinations for breast cancer risk prediction. *Pattern recognition*, 132:108919, 2022.
8. Hyeonsoo Lee, Junha Kim, Eunkyung Park, Minjeong Kim, Taesoo Kim, and Thijs Kooi. Enhancing breast cancer risk prediction by incorporating prior images. *arXiv preprint arXiv:2303.15699*, 2023.
9. Qianwei Zhou, Margarita Zuley, Yuan Guo, Lu Yang, Bronwyn Nair, Adrienne Vargo, Suzanne Ghannam, Dooman Arefan, and Shandong Wu. A machine and human reader study on ai diagnosis model safety under attacks of adversarial images. *Nature communications*, 12(1):7281, 2021.
10. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
11. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
12. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

12

13. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
14. Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.
15. Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.
16. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
17. Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.
18. Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018.
19. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
20. Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.
21. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
22. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
23. Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.
24. Jiafeng Wang, Zhaoyu Chen, Kaixun Jiang, Dingkang Yang, Lingyi Hong, Pinxue Guo, Haijing Guo, and Wenqiang Zhang. Boosting the transferability of adversarial attacks with global momentum initialization. *arXiv preprint arXiv:2211.11236*, 2022.
25. Chen Wan and Fangjun Huang. Adversarial attack based on prediction-correction. *arXiv preprint arXiv:2306.01809*, 2023.
26. Hegui Zhu, Yuchen Ren, Xiaoyan Sui, Lianping Yang, and Wuming Jiang. Boosting adversarial transferability via gradient relevance attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4741–4750, 2023.