scientific reports

Check for updates

OPEN

Mobile applications for skin cancer detection are vulnerable to physical camera-based adversarial attacks

Junsei Oda[1] & Kazuhiro Takemoto[1,2]✉

Skin cancer is one of the most prevalent malignant tumors, and early detection is crucial for patient prognosis, leading to the development of mobile applications as screening tools. Recent advances in deep neural networks (DNNs) have accelerated the deployment of DNN-based applications for automated skin cancer detection. While DNNs have demonstrated remarkable capabilities, they are known to be vulnerable to adversarial attacks, where carefully crafted perturbations can manipulate model predictions. The vulnerability of deployed medical mobile applications to such attacks remains largely unexplored under real-world conditions. Here, we investigate the susceptibility of three DNN-based medical mobile applications to physical adversarial attacks using transparent camera stickers under black-box conditions where internal model architectures are inaccessible. Through digital experiments with various DNN architectures trained on a publicly available skin lesion dataset, we first demonstrate that camera-based adversarial patterns can achieve high transferability across different models. Using these findings, we implement physical attacks by attaching optimized transparent stickers to mobile device cameras. Our results show that these attacks successfully manipulate application predictions, particularly for melanoma images, with attack success rates reaching 50–80% across all applications while maintaining visual imperceptibility. Notably, melanoma images showed consistently higher vulnerability compared to nevus images across all tested applications. To the best of our knowledge, this is the first demonstration of real-world adversarial vulnerabilities in deployed medical mobile applications, revealing significant security concerns where prediction manipulation could affect diagnostic processes. Our study demonstrates the importance of security evaluation in deploying such applications in clinical settings.

Keywords Deep neural networks, Medical imaging, Adversarial attacks, Security and privacy

Skin cancer ranks among the most prevalent malignant tumors worldwide. Early detection significantly improves patient prognosis[1–3]. However, the shortage of dermatologists, especially in resource-limited regions, creates critical gaps in diagnostic accessibility[4–6]. Since skin cancer develops in visually accessible areas and can be relatively easily evaluated through digital imaging, mobile applications utilizing device cameras have been developed to assess skin lesions[7–11]. These applications are expected to serve as primary screening tools, helping determine the necessity of specialist consultations. Conventional mobile applications, however, have faced scrutiny regarding their clinical utility and diagnostic accuracy[7,12–14].

Advances in artificial intelligence (AI), particularly deep learning technology[15], have transformed the landscape by significantly improving automated skin lesion classification[16–20], with some studies reporting performance comparable to specialists under specific conditions[21,22]. This technological progress has accelerated the development of mobile applications incorporating deep learning, marking a transition to a new era in automated skin cancer diagnosis[23–26].

Despite these advances, deep learning models introduce new security vulnerabilities. Of particular concern are adversarial attacks-techniques that deliberately manipulate neural network outputs by adding imperceptible perturbations to input data[27–29]. Research on adversarial vulnerabilities in deep learning-based medical image systems has been conducted across various modalities[30,31], including dermatological images[32,33], chest X-rays[33–35], optical coherence tomography images[33], and magnetic resonance imaging scans[36]. However, these studies have primarily focused on white-box analyses where model parameters are accessible, and have evaluated digital perturbations rather than physical attacks. In contrast, real-world mobile applications operate in black-

[1]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka, Japan. [2]Data Science and AI Research Center, Kyushu Institute of Technology, Iizuka, Fukuoka, Japan. ✉email: takemoto@bio.kyutech.ac.jp

box environments where model architecture is inaccessible and input data comes through physical camera capture.

This reality necessitates examining attack vulnerabilities under black-box conditions that reflect actual deployment environments, especially as these tools increasingly influence healthcare decisions and patient outcomes.

Evaluating the security of mobile applications used for medical diagnosis is crucial as these tools increasingly influence healthcare decisions and patient outcomes. Unlike controlled clinical environments, mobile applications operate in diverse settings where security vulnerabilities could affect large populations, potentially compromising patient safety and undermining public trust in digital health technologies[35,37]. The black-box nature of many deployed applications further emphasizes the need for robust external security evaluation frameworks.

Various physical adversarial attack methods have been proposed[38–40], including direct object manipulation (object attacks)[41,42], lighting environment manipulation (optical attacks)[43,44], and camera system intervention (camera-based attacks)[45]. Among these, camera-based attacks are particularly concerning for mobile applications as they can introduce relatively imperceptible perturbations across the entire captured image by simply attaching a transparent sticker with specific patterns to the camera lens. Since applying protective films to smartphone camera lenses is a common practice, this represents one of the most realistic attack scenarios against mobile applications.

However, previous study on camera-based attacks has been limited to white-box conditions[45], while black-box attack methods that leverage transferability have been widely studied in image recognition[46–48] and digital attacks on medical systems[49,50]. The integration of these approaches-combining physical camera-based attacks with black-box transferability against actual medical applications-remains uninvestigated, creating a critical gap in our understanding of real-world vulnerabilities.

In this study, we develop an attack framework that combines physical camera-based attacks with black-box transferability to evaluate the vulnerability of mobile applications for skin cancer detection. By quantitatively assessing the impact under realistic attack scenarios using only limited knowledge of model architecture, we aim to reveal novel vulnerabilities potentially inherent in medical diagnosis systems and provide insights for more secure system design.

Materials and methods

Target application and black-box environment

In September 2024, we searched for "skin cancer" on Google Play and Apple App Store, selecting three applications that explicitly stated the use of neural networks: DermoApp, SkiniveMD, and Blemish Types.

DermoApp[51], developed by Alien App Developer and downloaded over 10,000 times on Google Play, operates without user registration or cloud processing. The application provides binary classification specifically for melanoma detection (melanoma or not melanoma), reporting an accuracy of 96.8%. It processes images entirely on-device without requiring data storage or internet connectivity.

SkiniveMD[20], developed by Skinive B.V. and with more than 10,000 downloads on Google Play, is designed as a pre-diagnostic tool for healthcare professionals including dermatologists, therapists, and nurses The developer states that the application complies with Class I medical device regulations (CE MDD) and ISO 13485 quality management standards. The application can analyze over 50 different skin conditions, including various types of skin cancer (melanoma, basal cell carcinoma, squamous cell carcinoma) and precancerous conditions, using standard smartphone camera images for most conditions. For cancer risk assessment, however, the application specifically requires dermoscopic images.

Blemish Types[52], developed by Turion Development in collaboration with the Technical University of Teruel and downloaded more than 1,000 times on Google Play, employs neural networks trained for over 600 processing hours. The application reports an accuracy of 70.5% for classifying skin lesions into seven diagnostic categories, including melanoma and other common skin conditions. The application supports both standard and dermoscopic imaging modes.

All applications operate as black-box systems with only input-output access. While their descriptions suggest the use of mobile-optimized deep learning models, we did not use this information in our experimental design to maintain the integrity of our black-box evaluation approach, only reviewing technical details after completing all experiments.

Dataset and surrogate models

To develop black-box attacks against target applications, we employed a transferability-based approach. This method involves training several independent surrogate models to generate adversarial perturbations. By testing perturbations from different architectures, we could identify which designs produce more effective attacks against unknown target models.

Since the target applications focus on skin cancer detection, we utilized the International Skin Imaging Collaboration (ISIC) 2018 dataset[53,54] for training our surrogate models. This dataset, previously used in related studies[21,55,56], was chosen because its classification structure closely aligns with commercial skin cancer detection applications, ensuring our models are trained on data distributions similar to real-world usage.

The dataset consists of 10,015 RGB images classified into seven categories: melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis/Bowens disease (intraepithelial carcinoma; AKIEC), benign keratosis (solar lentigo/seborrheic keratosis/lichen planus-like keratosis; BKL), dermatofibroma (DF), and vascular lesions (VASC). Following our previous studies[33,55,56], we maintained the same data split with an approximate 70:30 ratio: 7,000 images for training (759 MEL, 4,679 NV, 358 BCC, 236 AKIEC, 790 BKL, 79 DF,

and 99 VASC) and 3,015 images for testing (354 MEL, 2,026 NV, 156 BCC, 91 AKIEC, 309 BKL, 36 DF, and 43 VASC).

For generating transferable adversarial perturbations, we carefully selected and trained multiple surrogate models that reflect potential architectures used in mobile applications. Our selection primarily focused on lightweight architectures suitable for mobile deployment: MobileNetV2[57], MobileNetV3-Large[57], ResNet-18[58], VGG-16[59], and EfficientNet-B1[60]. To comprehensively evaluate transferability, we also included their larger counterparts including ResNet-50[58], VGG-19[59], DenseNet-121[61], and DenseNet-201[61]. Additionally, considering the recent success of transformer architectures in computer vision tasks (vision transformer; ViT)[62], we incorporated ViT-Base-16 and ViT-Small-16 in our evaluation. This diverse set of architectures, ranging from mobile-optimized to more complex models, allows us to assess which type of model architecture generates perturbations that are most effective against mobile applications.

The training protocol followed the standard practices for utilizing pre-trained models. Input images were resized to 224×224 pixels to match the original training dimensions of the pre-trained models. During training, we applied data augmentation using random augmentation[63] with two operations at magnitude 14, and normalized the images (original pixel intensity scaled to 0–1), using mean and standard deviation values of 0.5 across all channels. The models were optimized using Adam[64] with an initial learning rate of 5.0×10^{-5}, incorporating a cosine learning rate schedule with warmup. To address the class imbalance in our dataset, we employed an imbalanced dataset sampler[65]. Mixed precision training was utilized to improve computational efficiency. We used a batch size of 256 and trained for 200 epochs, selecting the model with the best validation performance.

All model training and digital experiments (see "Digital experiments setup") were conducted using dual NVIDIA A100 GPUs (40GB each) with 192GB system RAM.

Physical camera-based attack framework

Our attack framework extends Li et al.'s[45] camera-based adversarial approach, adapting it for black-box medical imaging contexts[46–48]. The key innovation is combining transparent camera stickers with transferability techniques to generate universal adversarial perturbations (UAPs)[66] that can misclassify any input captured through the compromised camera. The implementation builds on Li et al.'s foundation with modifications for transferability across multiple surrogate models, parameter adjustments for medical imaging, and integration with the ISIC dataset pipeline.

Following previous work[45], we set the radius of each dot to 0.1 relative to the image size. Through empirical observation of actual transparent stickers attached to camera lenses, we determined two physical parameters: the opacity level ($\alpha = 0.2$) to reproduce the translucency of the sticker material, and the smoothness of the dropoff from the dot center ($\beta = 0.6$) to reflect the natural blending effect of the dots when viewed through the camera lens.

While these physical parameters remain constant, our framework optimizes two key parameters for each dot: its center position and RGB color values. We focus on non-targeted attacks because the target applications employ different classification schemes, making targeted attacks impractical in real-world scenarios. For generating the adversarial patterns, we employed a straightforward loss function based on cross-entropy[67]: $L = -\log(H(k(x_{\text{adv}}), y))$, where H represents the cross-entropy loss, $k(x_{\text{adv}})$ denotes the model's prediction for the perturbed image x_{adv}, and y is the true label. In practice, the loss is computed over mini-batches for computational efficiency. While more sophisticated loss functions have been proposed in previous work, we found that this simple formulation was effective for our physical attack scenario.

The optimization process employs the Adam optimizer with a cosine learning rate schedule, starting from an initial learning rate of 0.001. To ensure physical feasibility, we constrain all parameters to the range [0, 1] during optimization. The dot positions and colors are initialized randomly and gradually refined through the optimization process. For this adversarial pattern generation, we used a batch size of 32 and trained for 30 epochs, selecting the pattern that achieved the best validation performance.

Complete implementation details, including all parameters and configurations, are available in our public repository (https://github.com/kztakemoto/advStickersMed).

Digital experiments setup

Before physical implementation, we evaluated the transferability of adversarial patterns in a controlled digital environment.

For each surrogate model, we generate adversarial patterns and evaluate their effectiveness when transferred to all other models. To investigate the relationship between the number of dots and attack effectiveness, we test patterns with varying numbers of dots (5, 10, 15, 20, 25, and 30 dots).

We evaluate the performance using Attack Success Rate (ASR), defined as the proportion of cases where the model's prediction differs from the true label when the adversarial pattern is applied to the input image. Using the ISIC 2018 test dataset, we measure the ASR for each combination of source and target models. All images are processed following the same preprocessing steps used during model training.

Through this evaluation, we aim to determine both the optimal number of dots and identify which surrogate models generate the most transferable adversarial patterns. These findings inform our subsequent physical experiments against mobile applications.

Physical experiments setup

Building upon the findings from digital experiments, we created an evaluation dataset by randomly sampling 100 MEL and 100 NV images from the ISIC 2018 test set. We focused on MEL and NV images as they represent

the primary diagnostic distinction that skin cancer detection applications address: differentiating potentially malignant melanomas from benign nevi. These images were assigned sequential identifiers (MEL_001 to MEL_100 and NV_001 to NV_100) to ensure consistent selection across applications.

From this dataset, we used the first 100 MEL and 50 NV images for DermoApp, and the first 20 images of each class for SkiniveMD and Blemish Types. The applications varied in processing speed: DermoApp returned results within seconds, while SkiniveMD and Blemish Types required 2–3 minutes per image. To ensure evaluation consistency, we conducted up to three trials per image. This approach balanced thoroughness with resource constraints while maintaining statistical significance. We evaluated physical attack effectiveness using the same ASR metric as in digital experiments.

For physical implementation, we used different devices based on application compatibility. Most applications were tested on an Android tablet (JUSTSYSTEM ZJ-JS202) as they were available on Google Play. However, for SkiniveMD, due to version compatibility issues, we used an iPad Air (4 th generation, Model A2316) with the application installed through the Apple Store.

The adversarial dot patterns were printed on transparent glossy film (PLUS Corp. IT-324 F-C) using a standard inkjet printer (Canon TR153). The size of the dots was adjusted according to the lens diameter of the mobile devices' cameras. The transparent stickers were carefully aligned and attached to the camera lenses of the test devices. In preliminary tests, we observed that different transparent film materials produced similar attack success rates, suggesting the dot configurations are more critical than specific film properties.

Ethical considerations guided our experimental design. To avoid ethical concerns regarding human subjects while maintaining experimental rigor, we displayed the ISIC 2018 test images on a DELL U2718Q 27-inch monitor rather than conducting direct patient photography. This approach enabled comprehensive testing with a diverse set of skin lesion images while ensuring no additional human data collection was required. The monitor used to display the ISIC images was calibrated to standard settings with consistent brightness (250 cd/m^2) and color temperature (6500 K) before conducting experiments. To ensure consistent positioning between the mobile device cameras and the displayed images, we used a fixed mounting system that maintained a constant distance (approximately 15 cm) and perpendicular angle between the camera lens and the display surface. All experiments were conducted under standard indoor lighting conditions (uniform overhead LED lighting at approximately 500 lux, measured using a light meter at the experimental setup). When switching between devices (Android tablet and iPad), we maintained identical positioning and calibration protocols to ensure experimental consistency, with the only difference being the specific device used for image capture. Our study balanced the need for security research transparency with responsible disclosure, focusing on raising awareness of system vulnerabilities using commonly available materials and established techniques.

To ensure reliable evaluation of the applications' predictions, we conducted up to three trials for each image. The prediction label was determined based on these trials, with different criteria depending on the application's characteristics. In general, if any incorrect prediction was observed during the trials, we adopted that prediction as the final label. However, for SkiniveMD, which showed prediction instability leading to lower accuracy, we adopted the most frequently predicted class across multiple trials. In cases where multiple classes appeared with equal frequency, we selected the prediction with higher confidence as the final label.

Since these applications use different classification schemes from the ISIC 2018 dataset, we established a mapping between their prediction labels and the ISIC 2018 classes based on medical knowledge.

Results

To evaluate the vulnerability of mobile applications for skin cancer detection against physical camera-based adversarial attacks, we first conducted digital experiments to assess the transferability of adversarial patterns across different model architectures. We then performed physical experiments to verify the effectiveness of these attacks in real-world scenarios using actual mobile applications.

Digital experiments

To ensure reliability for generating adversarial patterns, we first evaluated our surrogate models on the ISIC 2018 test dataset. All models demonstrated robust performance with accuracies from 84.0% to 90.2%. Vision transformer architectures exhibited the strongest performance, with ViT-Base-16 and ViT-Small-16 achieving 90.2% and 89.6% accuracy respectively. Other architectures also performed well: ResNet-18 (86.8%), ResNet-50 (87.8%), DenseNet-121 (87.6%), and DenseNet-201 (88.0%). The VGG family showed similar performance levels with VGG-16 at 86.1% and VGG-19 at 85.8%. Notably, even lightweight architectures designed for mobile deployment maintained competitive performance, including MobileNetV2 (84.0%), MobileNetV3 (85.9%), and EfficientNet-B1 (85.6%). These results indicate that our surrogate models possess sufficient discriminative power for the skin lesion classification task and are suitable for generating transferable adversarial patterns.

To evaluate the transferability of adversarial patterns, we generated patterns using each surrogate model with 25 dots and tested them against all models. Table 1 shows the ASR for each combination of surrogate (rows) and target models (columns). For comparison, we also included results from random dot patterns (bottom row) where dot positions and colors were randomly determined, showing the average ASRs across 10 different random patterns.

The results demonstrate significant transferability of adversarial patterns across different architectures. For each surrogate model, we evaluated ASRs against all other models (excluding the surrogate itself). EfficientNet-B1 generated particularly effective patterns, achieving high ASRs (83.1–95.0%) against most targets, though effectiveness against ViT-Small-16 dropped to 47.9%. Similarly, ResNet-18 demonstrated strong transferability with ASRs consistently above 80% against most models, while falling to 44.3% against ViT-Small-16. In contrast, the ASRs of random patterns remained consistently low (15.3–25.4%), approximately equal to 1 minus the accuracy of each model, indicating that random patterns hardly contributed to misclassification. This clear

	A	B	C	D	E	F	G	H	I	J	K
A) ResNet-18	95.3	83.5	88.8	88.2	89.4	79.4	90.4	90.6	92.2	44.3	88.5
B) ResNet-50	82.0	85.8	65.2	61.1	70.2	62.2	69.1	68.0	84.6	31.7	75.5
C) VGG-16	84.4	74.4	90.8	85.4	81.0	68.2	82.7	80.2	82.4	27.2	68.9
D) VGG-19	74.7	76.4	83.3	88.3	81.6	61.4	79.4	68.7	74.8	28.5	60.3
E) MobileNetV2	74.1	80.7	79.9	81.7	90.7	82.3	87.1	75.9	82.7	31.4	61.3
F) MobileNetV3	50.2	45.2	46.2	44.6	49.2	49.5	48.4	42.4	47.7	29.2	42.0
G) EfficientNet-B1	88.2	86.3	91.2	94.3	93.3	89.1	95.0	83.1	87.8	47.9	71.3
H) DenseNet-121	73.9	77.2	56.8	52.5	62.2	51.9	67.9	89.7	77.8	32.3	76.7
I) DenseNet-201	82.3	79.8	72.9	70.6	74.1	56.1	77.3	66.7	92.5	31.4	79.4
J) ViT-Small-16	62.9	66.5	47.4	60.8	61.4	44.9	62.5	61.3	73.8	31.0	80.0
K) ViT-Base-16	64.1	46.3	34.8	40.7	40.5	31.7	46.8	49.2	57.9	29.3	90.9
Random	23.6	18.2	15.9	16.6	21.6	20.4	25.4	16.2	19.0	18.8	15.3

Table 1. Attack Success Rate (ASR; %) of adversarial patterns with 25 dots across different model architectures. Rows represent surrogate models used to generate adversarial patterns, while columns represent target models under attack. Letters A–K in column headers correspond to the same model architectures as indicated by the letters preceding the surrogate model names in rows (e.g., A: ResNet-18, B: ResNet-50, etc.). The "Random" row shows results from random patterns where dot positions and colors were randomly determined. Higher ASR indicates more successful attacks.

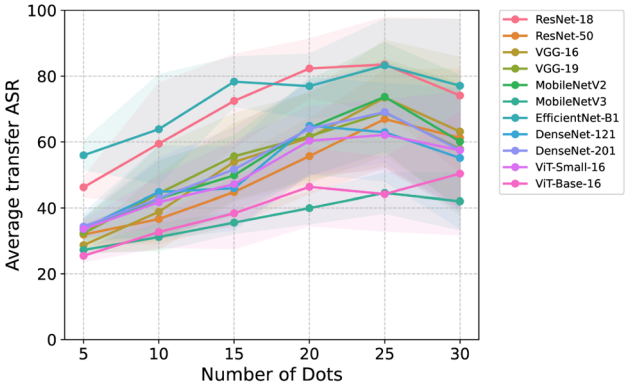

Fig. 1. Impact of the number of dots on adversarial pattern transferability. Average transfer Attach Success Rate (ASR; %) versus the number of dots for each surrogate model. Average transfer ASR is calculated as the mean ASR across all target models excluding the surrogate model itself. Shaded regions represent one standard deviation. Higher ASR indicates more successful attacks.

difference from optimized patterns confirms that the optimization of dot positions and colors is crucial for successful attacks.

Notably, Vision Transformer architectures (ViT-Small-16 and ViT-Base-16) showed relatively lower transferability compared to convolutional architectures, with ASRs generally ranging from 29.3% to 80.0%. This suggests that the architectural differences between transformers and convolutional networks may affect the transferability of adversarial patterns. Interestingly, patterns generated using MobileNetV3 showed limited transferability across all target models (29.2–50.2%), despite its competitive classification performance.

The diagonal elements of the table represent white-box scenarios where the target model is identical to the surrogate model. As expected, these cases generally showed higher ASRs (ranging from 49.5% to 95.3%, except for ViT-Small-16 at 31.0%), though the margin of improvement over cross-architecture transfer was not always substantial.

To understand why we focused on patterns with 25 dots in our transferability analysis (Table 1), we investigated how the number of dots affects the transferability of attacks. Fig 1 shows the relationship between the number of dots and the average transfer ASR for each surrogate model, where average transfer ASR is calculated as the mean ASR across all target models excluding the surrogate model itself. The shaded regions represent one standard deviation. Most models showed a generally increasing trend in ASR as the number of dots increased from 5 dots (average transfer ASR: 25–55%) to 25 dots (average transfer ASR: 45–85%). EfficientNet-B1 and ResNet-18 demonstrated particularly strong transferability, reaching average transfer ASRs of 83.3% and 83.5% respectively at 25 dots. Beyond 25 dots, the effectiveness either plateaued or decreased, as seen in these models where the average transfer ASR dropped to around 75% at 30 dots. While ViT architectures exhibited relatively lower transferability overall, they showed slightly different trends: ViT-Base-16 continued to show a marginal

increase up to 30 dots, though its maximum average transfer ASR remained relatively low at 50.4%. The overall trend across architectures indicated that 25 dots provided the optimal balance for attack transferability. This finding led us to focus our detailed cross-architecture transferability analysis on patterns with 25 dots, as presented in Table 1.

To understand misclassification patterns, we analyzed confusion matrices for selected models attacked using ResNet-18 as the surrogate model. We chose ResNet-18 for its strong transferability demonstrated in Table 1 (consistently achieving 79.4–92.2% ASR against various targets, except 44.3% against ViT-Small-16). Fig 2 shows row-normalized confusion matrices for architectures commonly used in mobile applications: VGG-16, MobileNetV3, and EfficientNet-B1.

A clear pattern emerges across all three models: adversarial patterns predominantly induced misclassification into the VASC (vascular lesions) class. For instance, when attacking VGG-16, 75.0–86.8% of samples from other classes were misclassified as VASC. Similar trends were observed in MobileNetV3 (30.8–58.3%) and EfficientNet-B1 (75.0–94.6%). This suggests that although our attack was non-targeted, the adversarial patterns consistently biased the models toward the VASC class.

Notably, NV (melanocytic nevus) images showed more resistance to the adversarial attacks compared to MEL (melanoma) images across all target models. For VGG-16, 13.9% of NV images were correctly classified compared to 4.5% of MEL images. Similar patterns were observed in MobileNetV3 (65.0% for NV vs 20.1% for MEL) and EfficientNet-B1 (13.9% for NV vs 4.5% for MEL). This resilience of NV images to adversarial attacks is particularly relevant for our subsequent physical experiments. Similar patterns of VASC-biased misclassification and NV resilience were observed across other surrogate-target model combinations not shown here, indicating that these characteristics are consistent features of our adversarial patterns rather than specific to particular model architectures.

Physical experiments

After confirming the effectiveness of camera-based adversarial attacks in digital experiments, we evaluated these attacks against actual mobile applications for skin cancer detection. First, we assessed each application's baseline performance without adversarial attacks as reference points. Since the applications use different classification schemes than ISIC 2018, we established a mapping between their prediction labels and ISIC classes based on medical knowledge (see Supplementary Table 1 for details). For DermoApp, we observed an accuracy of 75.3% (113/150 images correctly classified). This is notably lower than the 96.8% accuracy reported by the developer, though this difference may be attributed to our specific test conditions using displayed images rather than direct camera capture of actual lesions. SkiniveMD achieved an accuracy of 57.5% (23/40 images), while Blemish Types showed an accuracy of 85.0% (34/40 images). These results showed considerable variation from their reported performance levels of approximately 90% and 70.5% respectively, with SkiniveMD performing below and Blemish Types performing above their reported accuracies. However, these discrepancies should be interpreted with caution as our evaluation conditions differ from the original validation settings, particularly in terms of image acquisition methods and the specific subset of test images used.

Having established these baseline performances, we proceeded with the adversarial attacks. To physically implement the adversarial attacks, we printed the optimized dot patterns on transparent film and attached them to the camera lenses of the mobile devices. Fig 3 shows an example of the adversarial sticker pattern generated using ResNet-18 as the surrogate model (A) and its physical implementation on transparent film (B). While the pattern consists of 25 dots, some dots overlap due to the optimization of their positions, resulting in fewer visible dots in the final pattern. The actual images captured through a camera with the adversarial sticker (C) show subtle color shifts compared to clean images, yet maintain the overall visual characteristics of the skin lesions. While these example images were captured using an internal camera for illustration purposes (as most

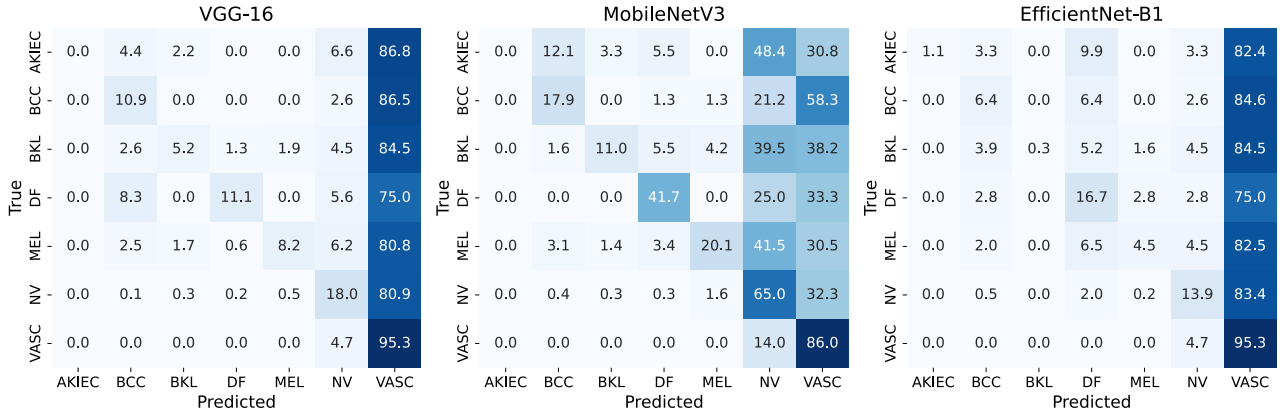

Fig. 2. Confusion matrices for selected target models under adversarial attack using ResNet-18 as surrogate model. Row-normalized confusion matrices for VGG-16, MobileNetV3, and EfficientNet-B1 under adversarial attacks. Each matrix element shows the percentage of images with the true label (row) that were classified as the predicted label (column). For the details of class labels (e.g, MEL and NV), see Dataset and surrogate models section.

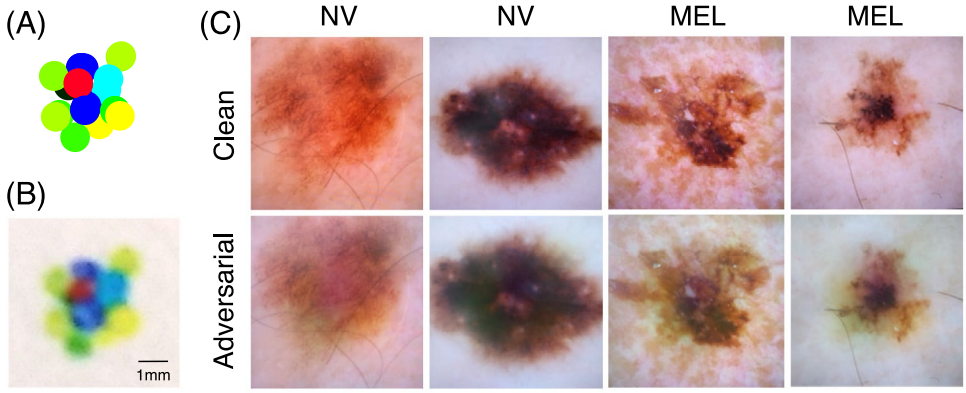

Fig. 3. Example of physical adversarial attack implementation. (**A**) Digital representation of the adversarial pattern (sticker) with 25 dots generated using ResNet-18 as the surrogate model. (**B**) Actual adversarial pattern printed on transparent film. (**C**) Example images of NV (melanocytic nevus) and MEL (melanoma) lesions captured without (Clean) and with (Adversarial) the sticker attached to the camera lens. Note that these images were captured using an internal camera for demonstration, as the target applications typically do not support image saving functionality. For detailed confidence scores of predictions before and after adversarial attacks, see Supplementary Table 1.

	DermoApp				SkiniveMD				Blemish Types			
	All		Correct		All		Correct		All		Correct	
	NV	MEL	NV	MEL	NV	MEL	NV	MEL	NV	MEL	NV	MEL
Surrogate models	(50)	(100)	(46)	(67)	(20)	(20)	(15)	(8)	(20)	(20)	(17)	(17)
ResNet-18	6.0	83.0	2.2	76.1	25.0	85.0	26.7	87.5	10.0	60.0	0.0	52.9
ResNet-50	16.0	79.0	10.9	76.1	25.0	80.0	20.0	75.0	10.0	50.0	5.9	41.2
VGG-16	10.0	85.0	6.5	79.1	25.0	80.0	20.0	75.0	10.0	50.0	5.9	41.2
VGG-19	12.0	76.0	8.7	67.2	10.0	80.0	6.7	75.0	15.0	65.0	5.9	58.8
MobileNetV2	16.0	81.0	13.0	73.1	15.0	75.0	6.7	62.5	5.0	50.0	0.0	41.2
MobileNetV3	26.0	73.0	23.9	64.2	15.0	70.0	6.7	50.0	10.0	60.0	5.9	58.8
EfficientNet-B1	12.0	77.0	6.5	65.7	15.0	80.0	0.0	75.0	15.0	70.0	11.8	64.7
DenseNet-121	22.0	70.0	17.4	59.7	15.0	80.0	6.7	62.5	25.0	65.0	17.6	58.8
DenseNet-201	14.0	77.0	8.7	67.2	15.0	80.0	0.0	75.0	5.0	65.0	0.0	58.8
ViT-Small-16	6.0	83.0	4.3	74.6	20.0	60.0	6.7	25.0	15.0	65.0	5.9	58.8
ViT-Base-16	8.0	86.0	6.5	80.6	20.0	60.0	6.7	50.0	15.0	60.0	5.9	52.9
Baseline	8.0	33.0	0.0	0.0	25.0	60.0	0.0	0.0	15.0	15.0	0.0	0.0

Table 2. Attack Success Rates (ASR; %) of physical adversarial camera-based attacks using 25-dot patterns against mobile applications. ASRs are shown separately for all images and images correctly classified at baseline. Numbers in parentheses indicate the number of test images in each category (NV: melanocytic nevus, MEL: melanoma). The "Baseline" row shows misclassification rates without adversarial attacks. Note that the baseline ASR for correctly classified images is 0.0% by definition. Higher ASR indicates more successful attacks.

applications do not allow saving captured images), they demonstrate how the adversarial sticker affects the image acquisition process with minimal visual disruption.

We evaluated attack effectiveness by applying camera stickers with 25 dots (optimal based on digital experiments) generated from different surrogate models. Table 2 shows ASRs for each surrogate-application combination. Supplementary Table 1 provides detailed image-by-image results, including confidence scores for predictions both with and without adversarial stickers across all tested applications.

Notably, ResNet-18, which showed the strongest transferability in digital experiments, also achieved the highest attack performance in physical settings. For DermoApp, it achieved ASRs of 83.0% for MEL images and 6.0% for NV images, compared to the baseline misclassification rates of 33.0% and 8.0% respectively. For SkiniveMD, ResNet-18 achieved an ASR of 85.0% for MEL images, while maintaining the baseline level ASR of 25.0% for NV images. For Blemish Types, it showed ASRs of 60.0% for MEL images and 10.0% for NV images, with the latter being lower than the baseline rate of 15.0%. EfficientNet-B1, which exhibited strong transferability in digital experiments, showed varied performance across applications: while achieving moderate ASRs for

DermoApp (77.0% for MEL and 12.0% for NV), it maintained relatively high effectiveness for SkiniveMD (80.0% for MEL and 15.0% for NV) and Blemish Types (70.0% for MEL and 15.0% for NV).

The lower ASR for NV images, particularly in SkiniveMD and Blemish Types, reveals an unexpected phenomenon: adversarial attacks sometimes improved classification accuracy by reclassifying previously misclassified images as NV (see Supplementary Table 1 for details).

To provide a more rigorous evaluation of attack effectiveness, we calculated ASRs specifically for images that were correctly classified without attacks. This analysis was particularly important for SkiniveMD, which showed a low baseline accuracy of 57.5%, and helps exclude cases where attacks appeared to improve NV classification by inadvertently correcting baseline misclassifications. This focused analysis revealed even clearer attack effectiveness: using ResNet-18-generated patterns, DermoApp showed ASRs of 76.1% for correctly classified MEL images (67 images) and only 2.2% for NV images (46 images). SkiniveMD showed more pronounced vulnerability, with ASRs of 87.5% for MEL images (8 images) and 26.7% for NV images (15 images) that were initially correctly classified. Blemish Types maintained similar resilience patterns, with ASRs of 52.9% for MEL (17 images) and 0.0% for NV (17 images) among correctly classified cases. Similar patterns were observed with other surrogate models: EfficientNet-B1 achieved ASRs of 65.7% and 6.5% for correctly classified MEL and NV images respectively in DermoApp, and ViT-Base-16 showed comparable effectiveness with ASRs of 80.6% and 6.5% for MEL and NV respectively. These results not only confirm the effectiveness of our attacks but also reinforce the observation that NV images consistently show more resilience to adversarial attacks compared to MEL images across all applications and surrogate models.

Discussion

Our study demonstrates that deep learning-based mobile applications for skin cancer detection are vulnerable to physical adversarial attacks under black-box conditions. Camera-based attacks using transparent stickers successfully manipulated predictions without requiring access to internal model architectures. To our knowledge, this is the first demonstration of such vulnerabilities in medical image diagnostic systems under real-world attack conditions. Previous studies[30–34,36] have primarily focused on digital perturbations under white-box conditions with fewer constraints that typically yield higher attack success rates (over 85% for UAPs in skin lesion classification[33]) while maintaining low perceptibility. In contrast, physical adversarial attacks generally achieve lower success rates due to environmental variables and physical limitations, as demonstrated in previous physical attack research[38,40]. What distinguishes our work is the achievement of substantial attack success rates (50–80%) despite the significant constraints of a black-box setting where model architecture is unknown and image manipulation is highly restricted. Moreover, compared to digital adversarial attacks, physical camera-based attacks present unique risks due to their persistence (affecting all images captured once applied), difficulty of detection (resembling standard protective films), and applicability across varying environmental conditions. The ability to successfully execute these attacks under such practical limitations, rather than ideal laboratory conditions, represents a significant finding with important security implications for deployed medical applications. Our findings extend beyond these theoretical vulnerabilities, showing that practical attacks are feasible against deployed medical applications. This raises significant concerns about the security of deep learning-based mobile medical applications, particularly as they become increasingly integrated into healthcare systems worldwide.

These vulnerabilities raise serious concerns about healthcare accessibility and equity[68]. In regions with limited access to dermatologists, where these applications serve as primary screening tools[4–6], manipulation could exacerbate healthcare disparities in two key ways. First, economically disadvantaged patients might manipulate diagnoses to avoid costly specialist consultations, potentially delaying critical melanoma detection. Second, false positives from manipulated images could burden healthcare systems, increase patient anxiety, and cause psychological distress even after eventual negative diagnosis. This is particularly critical for melanoma, where delayed detection can significantly impact survival rates[1–3].

These vulnerabilities could be exploited for various malicious purposes[32,33,35]. From an economic perspective, potential attacks range from insurance fraud through manipulated diagnoses to competitive sabotage among application developers. Individual motivations might include deliberate misdiagnosis for personal gain or revenge, while broader sociopolitical motivations could involve undermining public trust in medical AI technologies. For example, attackers could systematically manipulate applications to produce false positives, eroding trust in digital health technologies and disrupting healthcare delivery systems[69]. These risks are particularly concerning given the increasing integration of AI-based medical applications into formal healthcare protocols and their growing influence on medical decision-making processes.

Several defensive approaches could be implemented against these adversarial attacks[28–30,40,70], including adversarial training[30,71–73] and preprocessing methods to remove perturbations[74–76]. However, these defenses often reduce model accuracy on clean images[77]. This challenge is compounded because camera-based attacks using transparent stickers operate as low-frequency perturbations[78], which are particularly difficult to defend against. Even state-of-the-art mechanisms struggle to maintain effectiveness against low-frequency attacks while preserving performance on legitimate inputs[79–81].

Specific strategies for securing mobile diagnostic applications include: multi-angle image validation[82] ensuring consistency across capture perspectives, anomaly detection mechanisms[30,83] identifying pipeline manipulations, hardware-based solutions like specialized lens coatings, and ensemble approaches combining multiple architectures[84,85] to reduce adversarial transferability. Emerging technologies such as blockchain-based model verification could provide tamper-evident logging of imaging sessions and model behaviors, potentially helping to detect unusual patterns indicative of attacks[86]. Hardware-assisted security mechanisms, including those designed to counter adversarial example attacks, could improve the security of on-device processing by leveraging hardware-based detection and defense strategies[87]. These approaches present different trade-offs

between security, usability, and computational efficiency that must be carefully considered in medical application development.

As we conducted this study from an external security evaluation perspective without access to the applications' development environment, we could not implement and test these defensive measures. Nevertheless, given these identified vulnerabilities, a comprehensive security approach is necessary. Application developers should implement multiple layers of defense, including both technical measures and operational safeguards such as multi-angle image validation and anomaly detection systems[82]. Furthermore, these applications should be subject to rigorous regulatory oversight including security auditing processes, while users must be properly educated about their role as supplementary screening tools. Clinicians should approach these tools with appropriate caution, using them only as supplementary screening aids rather than diagnostic replacements, while remaining vigilant about their potential vulnerabilities to manipulation in real-world settings.

Regulatory frameworks for medical AI could be enhanced by incorporating adversarial security testing into approval processes, requiring transparency about security evaluation methodologies, and establishing post-market surveillance for emerging threats. These measures would help ensure that AI-based screening tools maintain their integrity against the types of vulnerabilities demonstrated in our study.

The security concerns are further complicated by the baseline performance of these applications. Although deep learning has demonstrated remarkable accuracy in controlled research settings[21,22], our results suggest that this performance may not directly translate to real-world mobile applications. This observation aligns with ongoing discussions regarding the clinical utility of mobile applications for skin cancer detection[7,12–14]. The integration of deep learning technology, while promising, not only shares similar accuracy limitations with conventional applications but also introduces new vulnerabilities that need careful consideration in clinical deployment.

While our digital experiments were primarily designed to inform the physical attacks, they also provide novel theoretical insights into adversarial vulnerabilities. Previous studies on camera-based adversarial attacks have been limited to white-box scenarios[45], but our results demonstrate that such attacks can be effectively executed under black-box conditions through transferability. This aligns with known properties of UAPs[66], as our camera sticker patterns essentially function as physical UAPs. The observed decrease in attack success rates with excessive numbers of dots can be attributed to physical constraints: as more dots are added to the limited space of the sticker, they increasingly overlap, reducing the effective degrees of freedom in the pattern and compromising the intended perturbation effects. Furthermore, the emergence of a dominant target class (VASC) is consistent with the characteristics of UAPs, where such dominant classes are commonly observed[33,66]. This dominance can be partially explained by the class imbalance in the training data: misclassifying images into minority classes like VASC can more effectively increase the overall attack success rate due to our definition of ASR. These findings extend our understanding of adversarial attacks beyond the digital domain, bridging the gap between theoretical vulnerabilities and practical attack scenarios.

Our study has important limitations that indicate directions for future research. While we demonstrated the feasibility of camera-based adversarial attacks, comprehensive clinical validation using real skin lesions is needed. Future studies should examine various environmental factors (lighting, camera angles) and patient demographics[88–90] to fully assess vulnerabilities in clinical settings.

The development and evaluation of defensive measures present another significant challenge. While we discussed potential defensive approaches like adversarial training and preprocessing methods, their practical implementation in mobile applications faces unique constraints, particularly in balancing security measures with computational efficiency. This challenge becomes more complex as mobile medical applications continue to evolve with new deep learning architectures and deployment strategies.

Beyond technical considerations, understanding the broader implications of security vulnerabilities in medical AI systems is crucial. Investigating how adversarial attacks interact with model interpretability and fairness across different demographic groups could contribute to developing more robust and equitable healthcare AI systems. This comprehensive approach to security research will be essential as AI-based medical applications become increasingly integrated into healthcare systems worldwide[37,91].

Conclusion

We demonstrated that deep learning-based mobile applications for skin cancer detection are vulnerable to physical adversarial attacks under black-box conditions, successfully manipulating predictions through transparent camera stickers particularly for melanoma images. These findings reveal significant security concerns in medical AI applications, where prediction manipulation could lead to delayed diagnosis or unnecessary medical interventions. While deep learning technology shows promise in medical image analysis, our results emphasize the need for careful consideration in deploying such applications in clinical settings, incorporating robust security measures and appropriate regulatory frameworks while ensuring their proper use as supplementary screening tools.

Data availability

The skin lesion images used in this study are publicly available from the ISIC 2018 dataset (https://challenge.isic-archive.com/data/#2018). The source code for implementing the adversarial attacks and reproducing our experiments is available at https://github.com/kztakemoto/advStickersMed. The mobile applications evaluated in this study (DermoApp, SkiniveMD, and Blemish Types) are available for download through Google Play Store and/or Apple App Store. Any additional information required to replicate this study is available from the corresponding author upon reasonable request.

Received: 31 January 2025; Accepted: 21 May 2025
Published online: 24 May 2025

References

1. Siegel, R.L., Giaquinto, A.N., & Jemal, A. Cancer statistics, 2024. CA: a cancer journal for clinicians. 74(1):12–49 (2024).
2. Diepgen, T. L. & Mahler, V. The epidemiology of skin cancer. *British Journal of Dermatology.* **146**(s61), 1–6 (2002).
3. Nikolaou, V. & Stratigos, A. Emerging trends in the epidemiology of melanoma. *British journal of dermatology.* **170**(1), 11–19 (2014).
4. Resneck, J. Jr. & Kimball, A. B. The dermatology workforce shortage. *Journal of the American Academy of Dermatology.* **50**(1), 50–54 (2004).
5. Ehrlich, A., Kostecki, J. & Olkaba, H. Trends in dermatology practices and the implications for the workforce. *Journal of the American Academy of Dermatology.* **77**(4), 746–752 (2017).
6. Porter, M. L. & Kimball, A. B. Predictions, surprises, and the future of the dermatology workforce. *JAMA dermatology.* **154**(11), 1253–1255 (2018).
7. Kassianos, A. P., Emery, J., Murchie, P. & Walter, F. M. Smartphone applications for melanoma detection by community, patient and generalist clinician users: a review. *British Journal of Dermatology.* **172**(6), 1507–1518 (2015).
8. Rat, C. et al. Use of smartphones for early detection of melanoma: systematic review. *Journal of medical Internet research.* **20**(4), e135 (2018).
9. Chan, S. et al. Machine learning in dermatology: current applications, opportunities, and limitations. *Dermatology and therapy.* **10**, 365–386 (2020).
10. Freeman, K., Dinnes, J., Chuchu, N., Takwoingi, Y., Bayliss, S.E., Matin, R.N., et al. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. bmj. 368. (2020)
11. Udrea, A. et al. Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. *Journal of the European Academy of Dermatology and Venereology.* **34**(3), 648–655 (2020).
12. Wolf, J. A. et al. Diagnostic inaccuracy of smartphone applications for melanoma detection. *JAMA dermatology.* **149**(4), 422–426 (2013).
13. Ngoo, A. et al. Efficacy of smartphone applications in high-risk pigmented lesions. *Australasian Journal of Dermatology.* **59**(3), e175–e182 (2018).
14. Deeks, J., Dinnes, J. & Williams, H. Sensitivity and specificity of SkinVision are likely to have been overestimated. *J Eur Acad Dermatol Venereol.* **34**(10), e582-3 (2020).
15. Shen, D., Wu, G. & Suk, H. I. Deep learning in medical image analysis. *Annual review of biomedical engineering.* **19**(1), 221–248 (2017).
16. Brinker, T. J. et al. Skin cancer classification using convolutional neural networks: systematic review. *Journal of medical Internet research.* **20**(10), e11936 (2018).
17. Goyal, M., Knackstedt, T., Yan, S. & Hassanpour, S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Computers in biology and medicine.* **127**, 104065 (2020).
18. Dascalu, A. & David, E. Skin cancer detection by deep learning and sound analysis algorithms: A prospective clinical study of an elementary dermoscope. *EBioMedicine.* **43**, 107–113 (2019).
19. Takiddin, A., Schneider, J., Yang, Y., Abd-Alrazaq, A. & Househ, M. Artificial intelligence for skin cancer detection: scoping review. *Journal of medical Internet research.* **23**(11), e22934 (2021).
20. Sokolov, K. & Shpudeiko, V. Dynamics of the neural network accuracy in the context of modernization of the algorithms of skin pathology recognition. *Indian Journal of Dermatology.* **67**(3), 312 (2022).
21. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., et al. Dermatologist-level classification of skin cancer with deep neural networks. nature. 542(7639):115–118. (2017)
22. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health.* **1**(6), e271–e297 (2019).
23. Kong, F. W., Horsham, C., Ngoo, A., Soyer, H. P. & Janda, M. Review of smartphone mobile applications for skin cancer detection: what are the changes in availability, functionality, and costs to users over time?. *International Journal of Dermatology.* **60**(3), 289–308 (2021).
24. Goceri, E. Diagnosis of skin diseases in the era of deep learning and mobile technology. *Computers in Biology and Medicine.* **134**, 104458 (2021).
25. Kränke, T. et al. New AI-algorithms on smartphones to detect skin cancer in a clinical setting-A validation study. *Plos one.* **18**(2), e0280670 (2023).
26. Smak Gregoor, A. M. et al. An artificial intelligence based app for skin cancer detection evaluated in a population based setting. *NPJ digital medicine.* **6**(1), 90 (2023).
27. Yuan, X., He, P., Zhu, Q. & Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems.* **30**(9), 2805–2824 (2019).
28. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. & Mukhopadhyay, D. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology.* **6**(1), 25–45 (2021).
29. Akhtar, N., Mian, A., Kardan, N. & Shah, M. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access.* **9**, 155161–155196 (2021).
30. Kaviani, S., Han, K. J. & Sohn, I. Adversarial attacks and defenses on AI in medical imaging informatics: A survey. *Expert Systems with Applications.* **198**, 116815 (2022).
31. Sorin, V., Soffer, S., Glicksberg, B.S., Barash, Y., Konen, E., & Klang, E. Adversarial attacks in radiology–A systematic review. European Journal of Radiology. p. 111085 (2023).
32. Finlayson, S. G. et al. Adversarial attacks on medical machine learning. *Science.* **363**(6433), 1287–1289 (2019).
33. Hirano, H., Minagi, A. & Takemoto, K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC medical imaging.* **21**, 1–13 (2021).
34. Asgari Taghanaki, S., Das, A., & Hamarneh, G. Vulnerability analysis of chest x-ray image classification against adversarial attacks. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, Proceedings 1. Springer; 2018. p. 87–94. (2018).
35. Hirano, H., Koga, K. & Takemoto, K. Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks. *Plos one.* **15**(12), e0243963 (2020).
36. Cheng, G. & Ji, H. Adversarial perturbation on MRI modalities in brain tumor segmentation. *IEEE Access.* **8**, 206009–206015 (2020).
37. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence.* **2**(6), 305–311 (2020).
38. Kurakin, A., Goodfellow, I.J., spsampsps Bengio, S. Adversarial examples in the physical world. In: Artificial intelligence safety and security. Chapman and Hall/CRC; p. 99–112 (2018).

39. Wei, H., Tang, H., Jia, X., Wang, Z., Yu, H., Li, Z., et al. Physical adversarial attack meets computer vision: A decade survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2024)
40. Ren, H., Huang, T. & Yan, H. Adversarial examples: attacks and defenses in the physical world. *International Journal of Machine Learning and Cybernetics.* **12**(11), 3325–3336 (2021).
41. Eykholt, K., & Evtimov, I. Fernandes E, Li B, Rahmati A, Xiao C, et al. Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition; p. 1625–1634 (2018).
42. Thys, S., Van Ranst, W., & Goedemé, T. Fooling automated surveillance cameras: adversarial patches to attack person detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops; p. 0–0 (2019).
43. Gnanasambandam, A., Sherman, A.M., & Chan, S.H. Optical adversarial attack. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; p. 92–101 (2021).
44. Wang, D., Yao, W., Jiang, T., Li, C., & Chen, X. Rfla: A stealthy reflected light adversarial attack in the physical world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; p. 4455–4465 (2023).
45. Li, J., Schmidt, F., & Kolter, Z. Adversarial camera stickers: A physical camera-based attack on deep learning systems. In: International conference on machine learning. PMLR; p. 3896–3904 (2019).
46. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., & Swami, A. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security; p. 506–519 (2017).
47. Cheng, S., Dong, Y., Pang, T., Su, H., & Zhu, J. Improving black-box adversarial attacks with a transfer-based prior. Advances in neural information processing systems. 32 (2019).
48. Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio B, Oprea, A, et al. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In: 28th USENIX security symposium (USENIX security 19); p. 321–338 (2019).
49. Bortsova, G. et al. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis.* **73**, 102141 (2021).
50. Ma, X. et al. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition.* **110**, 107332 (2021).
51. Kousis, I., Perikos, I., Hatzilygeroudis, I. & Virvou, M. Deep learning methods for accurate skin cancer recognition and mobile application. *Electronics.* **11**(9), 1294 (2022).
52. Turion Development. Blemish Types, Skin Cancer ID; . Available from: https://play.google.com/store/apps/details?id=com.FotoCheckSkin.(2024)
53. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data.* **5**(1), 1–9 (2018).
54. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368. (2019).
55. Minagi, A., Hirano, H. & Takemoto, K. Natural images allow universal adversarial attacks on medical image classification using deep neural networks with transfer learning. *Journal of Imaging.* **8**(2), 38 (2022).
56. Koga, K. & Takemoto, K. Simple black-box universal adversarial attacks on deep neural networks for medical image classification. *Algorithms.* **15**(5), 144 (2022).
57. Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. (2017).
58. He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; (2016). p. 770–778.
59. Simonyan, K. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. (2014)
60. Tan, M., & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR; p. 6105–6114 (2019).
61. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K.Q. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; p. 4700–4708 (2017).
62. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. (2020).
63. Cubuk, E.D., Zoph, B., Shlens, J., & Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops; p. 702–703 (2020).
64. Kingma, D.P. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. (2014).
65. Ming. Imbalanced Dataset Sampler; (2022). Available from: https://github.com/ufoym/imbalanced-dataset-sampler.
66. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., & Frossard, P. Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition; (2017). p. 1765–1773.
67. Poursaeed, O., Katsman, I., Gao, B., & Belongie, S. Generative adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition; p. 4422–4431 (2018).
68. Richardson, S., Lawrence, K., Schoenthaler, A. M. & Mann, D. A framework for digital health equity. *NPJ digital medicine.* **5**(1), 119 (2022).
69. Abernethy, A., Adams, L., Barrett, M., Bechtel, C., Brennan, P., Butte, A., et al. The promise of digital health: then, now, and the future. NAM perspectives. 2022;(2022)
70. Muoka, G. W. et al. A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense. *Mathematics.* **11**(20), 4272 (2023).
71. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In: International Conference on Learning Representations; (2018).Available from: https://openreview.net/forum?id=rJzIBfZAb.
72. Ma, L. & Liang, L. Increasing-margin adversarial (IMA) training to improve adversarial robustness of neural networks. *Computer methods and programs in biomedicine.* **240**, 107687 (2023).
73. Hu, L., Guo, X., Zhou, D., Wang, Z., Dai, L., Li, L., et al. Development and validation of a deep learning model to reduce the interference of rectal artifacts in MRI-based prostate cancer diagnosis. Radiology: Artificial Intelligence. 6(2):e230362 (2024).
74. Samangouei, P., Kabkab, M., & Chellappa, R. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. In: International Conference on Learning Representations; (2018).Available from: https://openreview.net/forum?id=BkJ3ibb0-.
75. Ashraf, S. N., Siddiqi, R. & Farooq, H. Auto encoder-based defense mechanism against popular adversarial attacks in deep learning. *PloS one.* **19**(10), e0307363 (2024).
76. Tran, K., Ly, L., & Luong, N.H. Adversarial Robustness of Medical Image Classifiers via Denoised Smoothing. In: International Symposium on Information and Communication Technology. Springer; (2024). p. 42–56.
77. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. Robustness May Be at Odds with Accuracy. In: International Conference on Learning Representations; (2019).Available from: https://openreview.net/forum?id=SyxAb30cY7.
78. Guo, C, Frank, J.S., & Weinberger, K.Q. Low Frequency Adversarial Perturbation. In: Adams RP, Gogate V, editors. Proceedings of The 35th Uncertainty in Artificial Intelligence Conference. vol. 115 of Proceedings of Machine Learning Research. PMLR; (2020). p. 1127–1137. Available from: https://proceedings.mlr.press/v115/guo20a.html.
79. Carlini, N., & Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proceedings of the 10th ACM workshop on artificial intelligence and security; (2017). p. 3–14.

80. Croce, F., & Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. PMLR; (2020). p. 2206–2216.
81. Chen, F. et al. Frequency constraint-based adversarial attack on deep neural networks for medical image classification. *Computers in biology and medicine.* **164**, 107248 (2023).
82. Aljedaani, B. et al. Challenges with developing secure mobile health applications: Systematic review. *JMIR mHealth and uHealth.* **9**(6), e15654 (2021).
83. Bulusu, S., Kailkhura, B., Li, B., Varshney, P. K. & Song, D. Anomalous example detection in deep learning: A survey. *IEEE Access.* **8**, 132330–132347 (2020).
84. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. Ensemble Adversarial Training: Attacks and Defenses. In: International Conference on Learning Representations; (2018).Available from: https://openreview.net/forum?id=rkZvSe-RZ.
85. Natha, S. et al. Automated brain tumor identification in biomedical radiology images: A multi-model ensemble deep learning approach. *Applied Sciences.* **14**(5), 2210 (2024).
86. Aliyu, I., Van Engelenburg, S., Mu'Azu, M. B., Kim, J. & Lim, C. G. Statistical detection of adversarial examples in blockchain-based federated forest in-vehicle network intrusion detection systems. *IEEE Access.* **10**, 109366–109384 (2022).
87. Zhang, J. et al. IEEE 29th Asian test symposium (ATS). *IEEE* **2020**, 1–6 (2020).
88. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences.* **117**(23), 12592–12594 (2020).
89. Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Mitigating bias in machine learning for medicine. *Communications medicine.* **1**(1), 25 (2021).
90. Glocker, B., Jones, C., Roschewitz, M., & Winzeck, S. Risk of bias in chest radiography deep learning foundation models. Radiology: Artificial Intelligence. 5(6):e230060 (2023).
91. Qayyum, A., Qadir, J., Bilal, M. & Al-Fuqaha, A. Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering.* **14**, 156–180 (2020).

Author contributions

K.T. conceived and designed the study. J.O. and K.T. jointly developed the attack framework. J.O. and K.T. designed the physical experiments, which J.O. conducted. Both authors contributed to the data analysis and interpretation of results. K.T. drafted the manuscript, and both authors reviewed, revised, and approved the final version.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-03546-y.

Correspondence and requests for materials should be addressed to K.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.