Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biomedical Engineering

Enhancing resilience against adversarial attacks in medical imaging using advanced feature transformation training

Danish Vasan[1] and Mohammad Hammoudeh[2]

This study presents a machine learning-driven defense mechanism against adversarial attacks, specifically tailored for medical imaging applications. This mechanism utilizes feature transformation through transfer learning, leveraging a fine-tuned ResNet152V2 network trained on original medical images. To enhance the model's robustness, we apply efficient adversarial training on transformed features extracted from both original and adversarial images. Additionally, we integrate Principal Component Analysis (PCA) to reduce feature dimensionality, optimizing the adversarial training process. When evaluated on Chest X-ray datasets, focusing on pneumonia and normal cases, the proposed mechanism demonstrated strong resilience against imperceptible attacks while maintaining a performance retention rate above 90 %. These results show the potential of the proposed mechanism to enhance the reliability and security of CNN-based medical imaging systems in practical, real-world settings.

Addresses
[1] Interdisciplinary Research Center for Intelligent Secure Systems, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia
[2] Department of Information and Computer Science, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

Corresponding author: Hammoudeh, Mohammad (m.hammoudeh@kfupm.edu.sa)

Current Opinion in Biomedical Engineering 2024, **32:**100561

This review comes from a themed issue on **Generalizable and Explainable Deep Learning for Medical Image Computing**

Edited by **Anand Paul**, **Naveen Chilamkurti** and **Awais Ahmad**

For complete overview of the section, please refer the article collection - Generalizable and Explainable Deep Learning for Medical Image Computing

Received 15 November 2023, revised 9 September 2024, accepted 27 September 2024

https://doi.org/10.1016/j.cobme.2024.100561

Keywords
Adversarial Attacks, Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Medical CNN Fine-tuning, Advanced Feature Transformation, Principal Component Analysis, Medical Imaging System.

Introduction

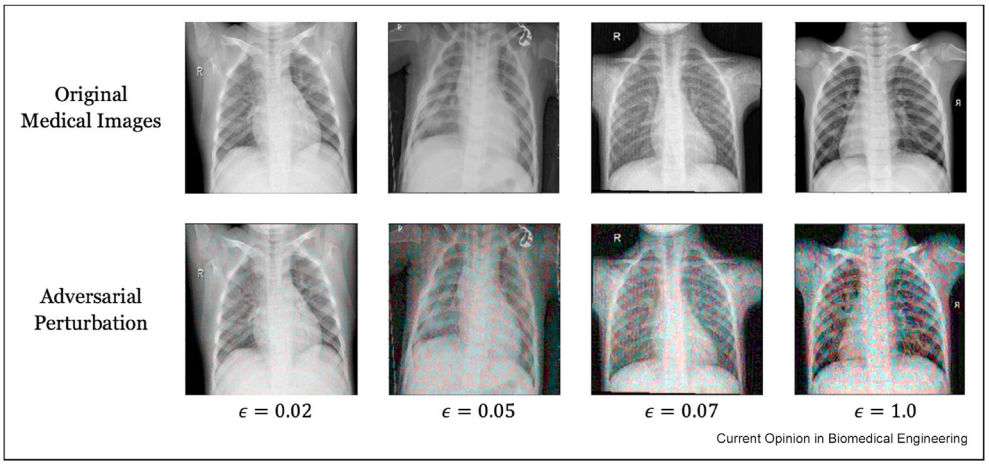
Medical imaging has a profound impact on healthcare by enabling the visualization of organs, cells, and pathological specimens, revolutionizing the diagnosis of diseases. It encompasses a broad range of tasks, including image processing, storage, analysis, retrieval, and interpretation. This technology plays a crucial role in advancing our understanding of various medical conditions, leading to more accurate diagnoses and improved treatment options.

Deep Neural Networks (DNNs), recognized as highly effective artificial intelligence techniques, gained significant prominence in medical imaging. They are essential in enhancing diagnostic accuracy and accelerating decision-making for healthcare professionals. DNNs achieved remarkable success in processing large volumes of data, extracting valuable insights from vast datasets across various fields. Examples of DNN applications in medical imaging include the early diagnosis of skin cancer through photographic images [1], classification of diabetic retinopathy using Optical Coherence Tomography Angiography (OCTA) images [2], pneumonia detection from chest X-rays [3], and nodule segmentation from CT images [4]. Various studies demonstrated that DNNs in medical imaging achieve performance levels comparable to human experts, with diagnostic results often matching those of professional medical staff.

Learning systems are proven highly cost-effective, delivering remarkable results with accuracy comparable to standard clinical practices. They received the United States Food and Drug Administration (FDA) approval. However, ensuring the security and reliability of medical DNNs remains a critical concern for the scientific community. Recent studies addressing classification and segmentation tasks in medical imaging revealed that even state-of-the-art DNNs are significantly susceptible to adversarial attacks. Medical imaging DNNs are more vulnerable to adversarial attacks than DNNs processing natural images as their input [5]. These vulnerabilities enable slight, carefully crafted manipulations in image samples, often invisible to the human eye, to significantly impact the DNN's performance (see Figure 1). Addressing these malicious adversaries is one of the most crucial challenges in medical Deep Learning (DL) systems.

Various methods were proposed to generate adversarial attacks, including the Fast Gradient Sign Method (FGSM) [6] and its more potent variants like Projected Gradient Descent (PGD) Madry2018, as well as the

Figure 1



The original medical image and adversarial image with different adversarial perturbations (ϵ).

"Carlini & Wagner (C&W)" method [7]. A study by Ref. [8] explained different incentives for attacking medical learning systems primarily from financial motivations. This is exacerbated by the rise in healthcare costs. Yet, medical DNNs cannot serve as replacements for physicians and medical specialists. Given the unique characteristics of medical images and the limited availability of annotated data, even with the guidance of human experts such as engineers, physicians, and radiologists, these neural networks were proven to be susceptible to rapidly evolving adversarial examples, highlighting the challenges of relying solely on automated systems for medical decision-making.

To counter adversarial attacks, researchers proposed various mitigation and detection techniques. One widely used approach involves adversarial training, incorporating adversarial samples into the training dataset to enhance the neural networks' robustness against adversarial attacks. Despite the effectiveness of these defense methods against adversarial attacks, they consistently show superior performance on CNNs with natural image datasets, such as CIFAR-10, compared to medical CNNs. This disparity in performance can be attributed to the insufficient availability of high-quality images and labeled data, proving the importance of addressing this data scarcity to strengthen the defenses of medical DNNs.

This study introduces a novel approach to adversarial medical machine-learning training focusing on robustness and efficiency. Unlike previous methods that augment adversarial samples into the training dataset, our approach emphasizes advanced feature transformation. Traditionally, augmenting adversarial samples demands significant computational resources and does not work well with the specific challenges posed by medical CNNs.

The primary contributions of this study include:

1. A review of existing research on adversarial medical CNN training, and analysis of state-of-the-art preprocessing methods to establish the foundation for the proposed approach.
2. Customizing and fine-tuning the advanced ImageNet model, specifically ResNet152V2, to better suit the unique characteristics of original medical images. This tailored approach enhances the network's performance in medical imaging contexts.
3. Designing white-box scenario dataset for FGSM attacks. This dataset is a robust testing ground for evaluating the resilience of the proposed approach under adversarial conditions.
4. Introducing an innovative feature transformation technique, leveraging transfer learning from the previously fine-tuned medical CNN. This approach enhances the model's ability to detect and counter adversarial attacks effectively.
5. Proposing an efficient and low computational cost strategy of adversarial training for medical networks without compromising the defense mechanisms.

The remainder of this article is structured as follows: Section Related work reviews related work on adversarial attacks, adversarial training, and data preprocessing methods. Section Proposed approach presents our proposed advanced feature transformation approach aimed at enhancing the robustness and efficiency of adversarial training. Section Use of PCA provides experimental

evidence supporting the effectiveness of the proposed method, while Section Evaluation and analysis concludes the article and outlines potential directions for future research.

Related work

This section reviews related work on adversarial attacks and defense mechanisms, including recent data preprocessing techniques and adversarial training methods aimed at improving the robustness of medical imaging systems.

Adversarial attacks

Since the introduction of adversarial examples by Ref. [9], numerous attack techniques have been proposed. Among these, gradient-based attacks which are classified as white-box attacks due to their access to complete model information were a central focus of exploration [10]. A novel approach in this field is the FGSM introduced in Ref. [6]. FGSM employs a gradient ascent technique using a first-order approximation of the loss function. Building on this, the author of [11] proposed the Basic Iterative Method (BIM), an extension of FGSM that applies iterative steps with a small step size, ensuring the perturbed image remains within the ϵ-neighborhood of the original through image clipping in each iteration. To further enhance BIM's performance and avoid local maxima, the Momentum-based Iterative Method (MIM) that incorporates a momentum term into the iterative attack process was introduced in "Boosting Adversarial Attacks with Momentum", 2018. Another significant variant is the PGD [12], which initializes perturbations within an l_∞ ball and projects the image back within that ball at each step. However, while these methods are effective, they do not necessarily find the minimal perturbation required to change the classification of a given input.

To address the issue of minimizing perturbations needed for misclassification, the author [13] introduced the Fast Adaptive Boundary Attack, which specifically aims to minimize the norm of the required perturbation. This approach provides a more precise method for generating adversarial examples. Another category of white-box attacks involves optimization-based methods. The C&W attack [7] uses a specialized objective function to generate adversarial examples. Additionally, the work [14] tackled adversarial attacks by framing sparse adversarial perturbations as a Mixed Integer Programming (MIP) problem, jointly optimizing perturbation magnitude and binary selection factors to enhance attack efficiency. Furthermore [13], introduced Auto Attack (AA), an advanced technique that extends the PGD attack. AA integrates two extensions of the PGD attack with two complementary methods, creating a powerful ensemble attack. This combination results in a highly effective method for challenging the robustness of machine learning models.

In this study, we focus on FGSM [6], which functions effectively in both targeted and untargeted attack scenarios. The primary goal is to regulate the l_1, l_2, or l_∞ norm of the adversarial perturbation. In the targeted case, particularly concerning the norm, the adversarial perturbation generated by the FGSM attack is formulated in Equation (1).

$$\Psi(x,y) = -\epsilon \cdot sign(\nabla_x \mathcal{L}(x,y)) \tag{1}$$

where $\epsilon > 0$ represents the attack strength and y denotes the target class specified by the attacker. The resulting adversarial sample is then calculated in Equation (2).

$$x_{adv} = clip(x + \Psi(x,y), x_{min}, x_{max}) \tag{2}$$

Intuitively, the attack transforms the input x to minimize the classifier's loss when classifying it as y. For the l_p norms with $p = 1, 2$, the adversarial perturbation is calculated in Equation (3).

$$\Psi(x,y) = \epsilon \cdot \left(\frac{\nabla_x \mathcal{L}(x,y)}{\|\nabla_x \mathcal{L}(x,y)\|_p} \right) \tag{3}$$

Adversarial defenses

Considering the harmful impact of adversarial examples on real-world applications, a wide range of strategies were developed to mitigate this issue and strengthen the robustness of models against adversarial attacks. Most existing defense methods can be broadly classified into two categories: data preprocessing and adversarial training.

Data preprocessing approaches

Several researchers have proposed data preprocessing techniques to protect neural networks from adversarial attacks during both training and testing phases. For instance, in [15], the author recommends using JPEG compression as an effective preprocessing step within the classification pipeline. This technique mitigates adversarial attacks by removing high-frequency components from images, effectively blurring out adversarial perturbations embedded in the image blocks. In [16], the author introduces the Gaussian Data Augmentation method, a well-established computer vision technique that enhances model robustness. This method augments the dataset by generating variations of the original samples with added Gaussian noise, helping the model become more resistant to adversarial perturbations. In [17], the author presents the Total Variance Minimization method, which involves selecting a small subset of pixels and reconstructing the simplest possible version of the image from these pixels. The reconstructed image is free of adversarial perturbations, as these perturbations are typically small and localized.

The author [18] proposed Spatial Smoothing, specifically for images as an adversarial defense method. This approach filters out adversarial signals by employing local spatial smoothing techniques. The author [19] introduced the Feature Squeezing technique, which transforms input data by reducing the precision of its components. In the case of images, this involves reducing the typical 8-bit pixel values to a lower number of bits, b, where b is less than 8, effectively reducing the potential space for adversarial perturbations.

The author [20] developed a preprocessing method known as thermometer encoding, designed for input discretization. This approach encodes each feature as a binary vector of fixed length by partitioning the input domain into b distinct buckets, where b represents the number of bits used to encode each feature. The number of ones in the binary vector corresponds to the index of the bucket containing the feature's original value, with the remaining positions filled with zeros. Thermometer encoding effectively preserves the pairwise ordering of features, helping to defend against adversarial manipulation. Unlike the previous state-of-the-art preprocessing approaches, our method leverages advanced feature transformation using a fine-tuned CNN specifically designed for medical imaging applications.

Adversarial training

Adversarial training is designed to improve the robustness of classifiers $C(x)$ by incorporating adversarial samples into the training set [6]. A specific variation of this approach, known as virtual adversarial training, generates adversarial samples using the Virtual Adversarial Method [21]. In Miyato et al., 2019, a combination of attacks and classifiers $(\rho_1, C_1), \ldots, (\rho_m, C_m)$ is utilized alongside the original training dataset $(x_1, y_1), \ldots, (x_n, y_n)$. Various tools, such as ART [22], ADVBOX [23], Cleverhans [24], FoolBox [25], and DEEPSEC [26], support adversarial training by augmenting the training set with adversarial samples. These tools statically add adversarial examples $(\rho_i(x_j), C_i(\rho_i(x_j)))$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. As a result, when n attacks are applied to all original training samples, the augmented dataset grows to a size of $m \times n$, which is then used to train a more robust classifier $C(x)$. Adversarial training is framed as an optimization problem as shown in Equation (4). The goal is to approximate the inner maximization problem by generating adversarial examples and subsequently updating the model parameters, denoted as θ.

$$\min_\theta E_{(x,y)\sim(x,y)}\left[\max_{\|X'-x\|_\infty < \epsilon} L(f_\theta(X'), y)\right] \tag{4}$$

Unlike conventional adversarial training methods, our approach trains the network using a custom dataset derived from high-dimensional features extracted by a fine-tuned medical CNN. Importantly, our method does not rely on data augmentation or additional network training.

Proposed approach

This section presents an overview of the datasets used in this study, followed by a brief explanation of the fine-tuning process applied to the ImageNet CNN using the original medical images. We then detail the advanced feature transformation approach employed to create a custom dataset for training adversarial medical networks. Next, we apply Principal Component Analysis (PCA) to eliminate redundant features and reduce the dimensionality of the feature vector. Finally, the medical network is trained on the refined feature vector as part of the adversarial training process. Figure 2 illustrates the workflow of the proposed approach.

Datasets

We utilized a medical dataset of X-ray images, consisting of 5216 samples for training and 624 samples for testing. The images were categorized into two classes: Normal and Pneumonia. To ensure compatibility with our CNN architecture, which requires fixed-size inputs, all images were standardized to a uniform size of 224×224 pixels. After this preprocessing step, various adversarial datasets were generated from the original medical dataset, each corresponding to a different adversarial perturbation ϵ. For the generation of these datasets under white-box attack conditions, we employed the FGSM using a fine-tuned medical CNN. The following subsections provide detailed descriptions of the fine-tuned CNN architecture, along with a comprehensive summary of both the original and adversarial datasets and their respective perturbation values, as outlined in Table 1.

Medical CNN fine-tuning

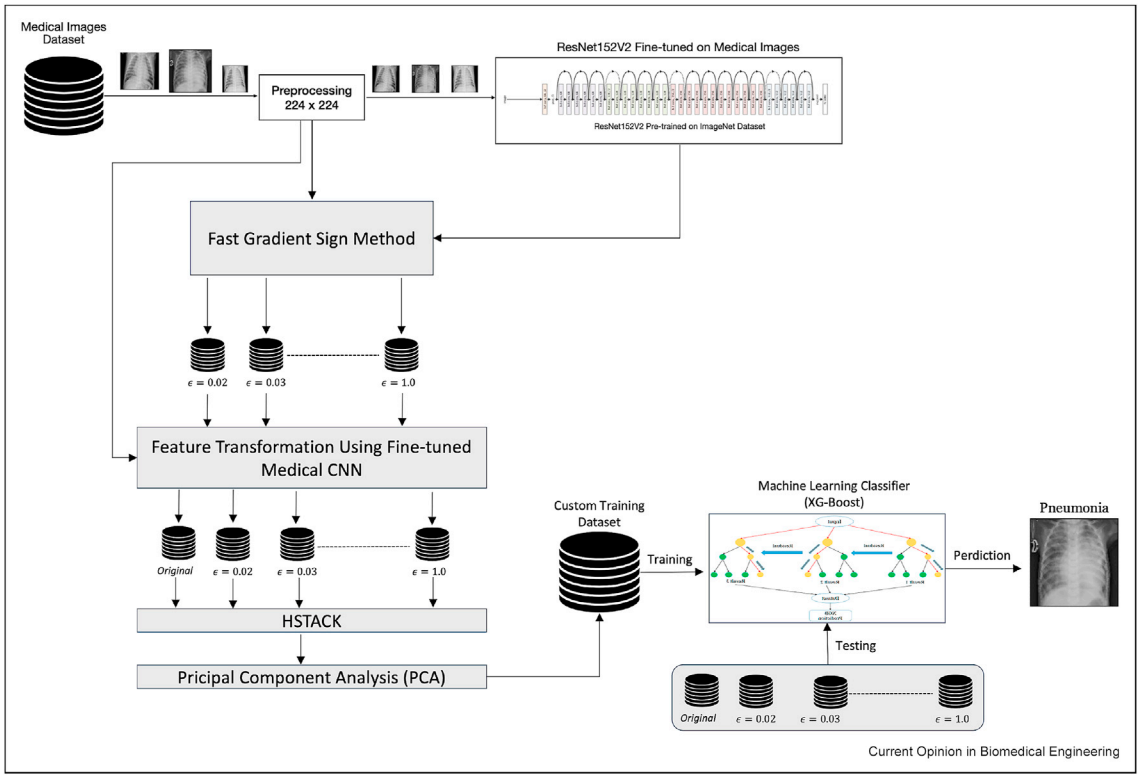
We adopted the ResNet152V2 architecture, originally trained on the ImageNet dataset [27], which consists of 1000 classes. To tailor the CNN to our specific problem, we replaced the final fully connected layer initially designed for 1000 classes with a new layer containing 2 classes (Normal and Pneumonia). Additionally, we introduced a weight regularization mechanism ($kernel\_regularizer = 0.0001$) to each layer, which helped mitigate the risk of overfitting during fine-tuning. The weights of the pre-trained CNN, initially optimized for natural images, were initialized and further refined using fine-tuning techniques via back-propagation.

Let X represent the training dataset consisting of n X-ray images. The fine-tuning process iteratively optimizes the filter weights w to minimize the error rate, as described by the optimization procedure in Equation (5).

$$L(w, X) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i, w), \widehat{c_i}) \tag{5}$$

where, $f(x_i, w)$ represents the CNN function used to predict the class c_i of input image x_i based on the parameters w. The term $\widehat{c_i}$ denotes the true class of the ith image x_i in the

Figure 2



The process of adversarial ML training.

dataset X. The penalty function $l(c_i, \widehat{c_i})$ quantifies the cost incurred when the predicted class c_i differs from the true class $\widehat{c_i}$. Typically, this penalty function l is implemented as a logistic loss function.

Feature transformation

After fine-tuning the ImageNet CNN with the original medical images, the model is used to perform feature transformation. Each sample is passed through the fine-tuned CNN to extract bottleneck features, resulting in 2048 features per sample. This process is applied to all samples in the datasets, encompassing both the original and adversarial samples. As a result, we collected 5216 training samples, each containing 22,528 features, and 624 testing samples, each also represented by 22,528 features. In the next step, we employ PCA to reduce redundancy in the feature vectors and lower the dimensionality of the features.

Let X represent the dataset containing N samples, each denoted by x_i, where $i = 1, 2, \ldots, N$. Each x_i is processed by a pre-trained model M, which extracts bottleneck features, represented as f_i. The process of extracting these bottleneck features can be described as a function F applied to each sample x_i (See Equation (6)).

$$f_i = F(x_i, M) \tag{6}$$

The bottleneck features f_i from all samples are stacked vertically to form a matrix F of size $N \times D$, where D represents the dimensionality of the bottleneck features shown in Equation (7).

$$F = \begin{pmatrix} f_1^T \\ f_2^T \\ \vdots \\ f_N^T \end{pmatrix} \tag{7}$$

Table 1

Statistical summary of the used datasets.

Name	Samples	Adversarial Perturbation (ϵ)
Chest X-Ray	5840	Original
Chest X-Ray	5840	0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 1.0

where f_i^T denotes the transpose of the bottleneck feature vector f_i. The stacking operation is denoted as Stack$(f_1, f_2, \ldots, f_N)$. Thus, the entire process can be calculated in Equation (8).

$$F = \text{hstack}(F(x_1, M), F(x_2, M), \ldots, F(x_N, M)) \tag{8}$$

Use of PCA

PCA is a widely used statistical technique for reducing dimensionality and visualizing data. In machine learning and data analysis, PCA simplifies complex datasets by transforming high-dimensional data into a lower-dimensional form while preserving essential information. After the feature transformation, PCA is applied to further reduce the size of the feature vector for each sample. In our case, this process reduces the number of features from 22,528 to 624 per sample. Figures 3 and 4 illustrate the visualization of the features before and after transformation, respectively.

Medical network adversarial training

Adversarial training is a technique designed to enhance the robustness of a classifier $C(x)$ by incorporating adversarial samples into the training dataset. In contrast to traditional adversarial training methods, our approach involves training the machine learning network with a custom dataset composed of transformed features extracted from a fine-tuned CNN architecture. Importantly, this custom dataset maintains the same number of training and testing samples and includes uncorrelated features for each sample. The inclusion of these uncorrelated features serves two key purposes: it accelerates the training process for the machine learning classifier and promotes robust learning, minimizing the risk of overfitting.

For adversarial training, we employed the XGBoost machine learning classifier due to its remarkable robustness, effective management of overfitting, and ability to provide valuable insights into feature importance. The training process for the XGBoost classifier using the transformed features dataset is described in Equation (9):

$$C_{Adv} = XGBoost(X_{PCA}, Y, \Theta) \tag{9}$$

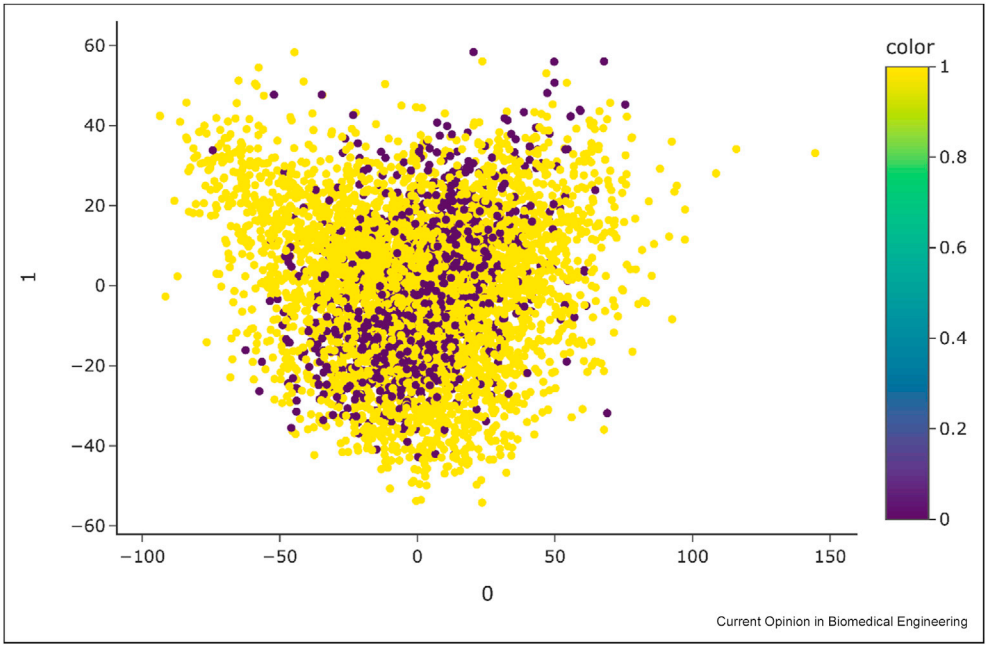
where C_{Adv} represents the robust adversarial classifier trained on the transformed features dataset, X_{PCA} denotes the input feature set, Y refers to the corresponding labels, and Θ includes the hyperparameters and configuration settings of the XGBoost model.

Evaluation and analysis

Experimental setup

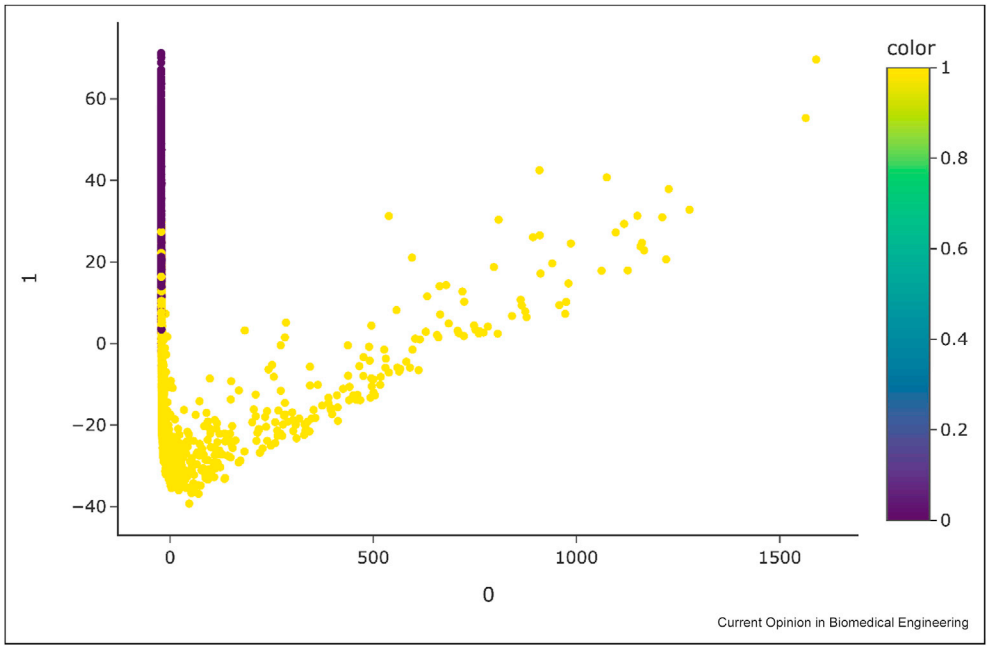
In this section, we evaluate the performance of the fine-tuned CNN on both the original and adversarial datasets. We implement adversarial training techniques through the ART library (Nicolae et al., n.d.) To further train the fine-tuned CNN, assessing its performance across both datasets. Finally, we train a machine learning classifier using our proposed feature transformation approach and evaluate its effectiveness on both the original and adversarial datasets.

Figure 3



Feature visualization before transformation.

Figure 4



Feature visualization after transformation.

To assess the performance of the proposed methods, we utilized several evaluation metrics, including accuracy, F1-score, recall, precision, and the Receiver Operating Characteristic (ROC) curve. These metrics are widely recognized in the research community. True Positive (TP) refers to instances where both the actual and predicted categories are positive. True Negative (TN) occurs when both the actual and predicted categories are negative. False Positive (FP) represents cases where the actual category is negative, but the predicted category is positive. False Negative (FN) indicates instances where the actual category is positive, but the predicted category is negative.

The experimental setup was developed in Python and executed on a Mac Studio Apple M1 Ultra system, equipped with a 48-core GPU, a 20-core CPU (comprising 16 performance cores and 4 efficiency

Table 2

Comparative summary.

Medical Images Datasets	Fin-tuned Medical CNN Accuracy (%)	Adversarial Training of Fine-tuned CNN Accuracy (%)	Proposed Adversarial Training Using Transformed Features Accuracy (%)
Original	91.02	77.53	86.38
$\epsilon = 0.02$	0.09	75.12	90.54
$\epsilon = 0.03$	0.11	75.01	90.87
$\epsilon = 0.04$	0.19	76.51	90.22
$\epsilon = 0.05$	0.26	77.19	88.30
$\epsilon = 0.06$	0.31	81.23	87.34
$\epsilon = 0.07$	0.32	75.59	85.74
$\epsilon = 0.08$	0.34	72.10	85.42
$\epsilon = 0.09$	0.33	75.41	84.62
$\epsilon = 0.10$	0.31	71.24	83.93

We calculated the computational cost for each experiment.
(1) Fine-tuned Medical CNN took 85 min.
(2) Adversarial Network Training took 403 h, and
(3) Proposed adversarial training took 5–10 min including feature transformation, reduction, & adversarial network training.

cores), and 64 GB of RAM. This system was used for both training and testing purposes.

Experiments on fine-tuned CNN

Fine-tuning a pre-trained CNN for specific tasks is a common practice in DL. Models pre-trained on large, diverse datasets like ImageNet capture general features that are useful for a variety of computer vision applications. In this experiment, we utilized a customized pre-trained CNN architecture, specifically ResNet152V2, which was originally trained on a dataset of natural images. We further fine-tuned this model using medical images. This fine-tuning process resulted in an accuracy of 91.02 % when tested on our medical dataset. Table 2 summarizes the fine-tuned medical CNN's performance across different levels of adversarial perturbations.

Experiments on adversarial trained CNN

In the next phase of our study, we applied traditional adversarial training to the previously fine-tuned medical CNN, using the Adversarial Robustness Toolbox (ART). Each original sample was augmented with various adversarial attacks, and the network was trained accordingly. It is important to note that the adversarial training process was computationally intensive, taking a total of 403 h to complete.

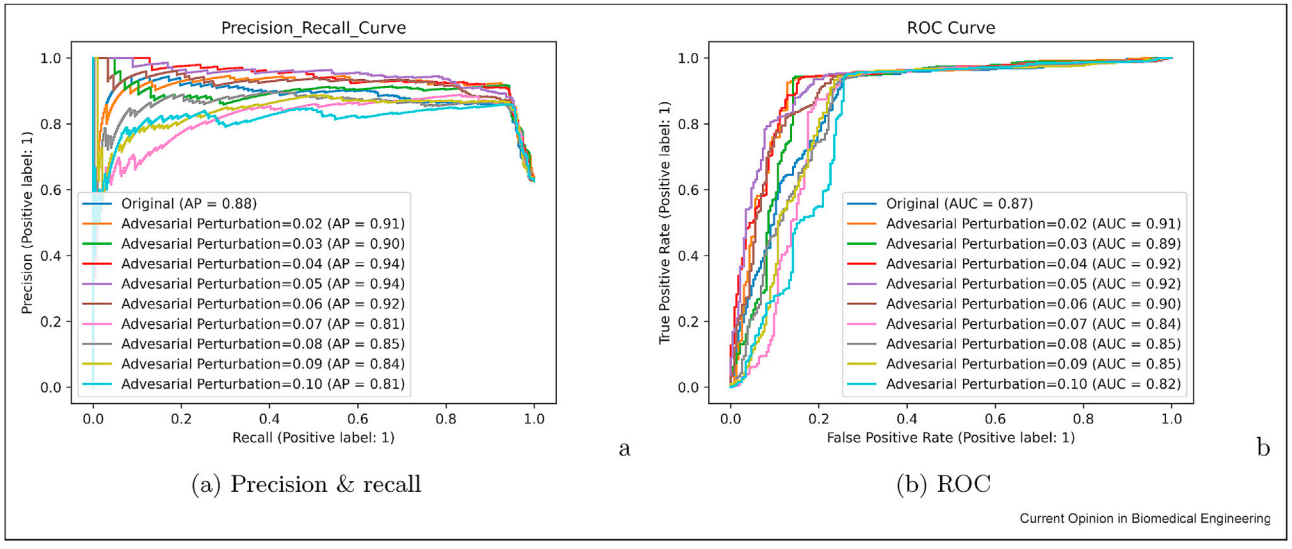
The results of this experiment showed a test accuracy of 77.53 % on the original dataset. However, when evaluated on the adversarial datasets, the accuracy ranged from 81.23 % to 71.24 % (See Table 2). A notable observation was that adversarial training led to a decrease in the model's classification performance on the original images, likely due to overfitting during the training process. This highlights the challenge of balancing the model's defense against adversarial attacks while maintaining accuracy on the original data.

Experiments on transformed features

Feature transformation is important in enhancing the accuracy and reliability of machine learning models. By effectively managing outliers and accommodating the inherent non-linearities present in real-world datasets, these transformations enable models to better capture the complexities of the data, resulting in more precise and insightful predictions. In this experiment, we evaluated a machine learning classifier trained on a custom dataset containing transformed features extracted from the fine-tuned medical CNN. This approach achieved an accuracy of 90.54 % on the original medical dataset. Notably, the model's accuracy ranged from 90.54 % to 81.09 % when tested on various adversarial datasets. Detailed results from this experiment with transformed features are presented in Table 2.

Figure 5 presents the precision, recall, and ROC curves of our proposed approach. The experimental results show the superiority of our method across multiple performance metrics. Additionally, we compared the training time of the machine learning classifier trained on transformed features with that of the fine-tuned CNN and traditional adversarial training. The findings indicate that adversarial training with feature transformation is both more robust and efficient compared to conventional adversarial training and fine-tuning approaches (See Table 2 footnote).

Figure 5



Experimental results.

Conclusion

This study addresses the critical issue of adversarial attacks on CNNs used in medical imaging, a challenge that significantly impacts the deployment of CNN-based solutions in clinical practice. To address this, we introduced a novel defense strategy that involves feature transformation via transfer learning from a fine-tuned ResNet152V2 network applied to the original medical images. This is followed by efficient adversarial training using transformed feature vectors from both original and adversarial medical images. Additionally, PCA was applied to effectively reduce the feature dimensionality. This approach was evaluated using a Chest X-ray image dataset to assess its resilience against white-box attacks. The experimental results demonstrate the effectiveness of the proposed approach, successfully defending against imperceptible adversarial attacks with a performance retention rate exceeding 90 %. These findings have important implications for enhancing the security and robustness of CNN-based medical applications in real-world settings, offering potential improvements in the reliability of medical imaging solutions.

Declaration of competing interest

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

Data availability

The authors made use of public datasets which are all reference in the manuscript.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest

1. * Mira ES, Sapri AMS, Aljehanõ RF, Jambõ BS, Bashir T, El-Kenawy E-SM, Saber M, *et al.*: **Early diagnosis of oral cancer using image processing and artificial intelligence**. *Fusion: Practice and Applications* 2024, **14**:293–308.

This study introduces a deep learning-based approach for diagnosing oral diseases using smartphone photographs. It presents a "centered rule" method for capturing highquality images and a resampling technique to address variability from handheld cameras. The approach aims to enhance oral cancer diagnosis through improved image capture and processing, demonstrating the potential of smartphone-based diagnostic tools.

2. * Bidwai P, Gite S, Pradhan B, Gupta H, Alamri A: **Harnessing deep learning for detection of diabetic retinopathy in geriatric group using optical coherence tomography angiography-octa: a promising approach**. *MethodsX* 2024:102910.

This study explores using Optical Coherence Tomography Angiography (OCTA) and Deep Learning to detect diabetic retinopathy (DR) in elderly patients. With a dataset of 262 OCTA scans, the authors trained various Convolutional Neural Networks (CNNs) to classify DR severity. The findings highlight the method's potential for enhancing early DR detection and supporting clinicians in managing age-related ocular diseases.

3. * Nalluri S, Sasikala R: **Pneumonia screening on chest x-rays with optimized ensemble model**. *Expert Syst Appl* 2024, **242**, 122705.

This study addresses the challenge of accurately diagnosing pneumonia from chest X-ray images by proposing a multi-stage model. The model includes pre-processing with Dynamic Histogram Equalization and median filtering, followed by Enhanced Watershed Segmentation for separating the region of interest. Feature extraction involves various pattern-based techniques, and the Archimedes-assisted Henry Gas Optimization Algorithm (AHGOA) is used for optimal feature selection. The approach aims to enhance diagnostic accuracy by effectively processing and analyzing chest X-ray images.

4. * Agnes SA, Solomon AA, Karthick K: **Wavelet u-net++ for accurate lung nodule segmentation in ct scans: improving early detection and diagnosis of lung cancer**. *Biomed Signal Process Control* 2024, **87**, 105509.

This paper presents Wavelet U-Net++, a novel approach for accurate segmentation of lung nodules in CT scans. The method integrates the U-Net++ architecture with Haar wavelet pooling to capture both high- and low-frequency image information, enhancing segmentation accuracy. The approach is evaluated on the LIDC-IDRI dataset and demonstrates improved performance over existing methods, particularly in detecting small and irregular nodules. The combination of wavelet pooling and advanced loss functions showcases a significant advancement in lung nodule segmentation.

5. * Ma X, Niu Y, Gu L, Wang Y, Zhao Y, Bailey J, Lu F: **Understanding adversarial attacks on deep learning based medical image analysis systems**. *Pattern Recogn* 2021, **110**, 107332.

This paper examines the vulnerability of deep neural networks (DNNs) used in medical image analysis to adversarial attacks. It reveals that medical DNN models are particularly susceptible to such attacks but also highlights that these attacks are often detectable due to distinct feature dicerences. The findings suggest that medical adversarial attacks can be identified with high accuracy, which could inform the development of more secure and explainable medical deep learning systems.

6. * Goodfellow IJ, Shlens J, Szegedy C: **Explaining and harnessing adversarial examples**. In *3rd international conference on learning representations, ICLR 2015 - conference track proceedings*; 2015:1–11.

This paper investigates the vulnerability of neural networks to adversarial examples, attributing their susceptibility to the networks' linear characteristics rather than nonlinearity or overfitting. The study presents new quantitative results supporting this view and introduces a simple method for generating adversarial examples. This method improves adversarial training, leading to reduced test set error on the MNIST dataset for a maxout network.

7. * Carlini N, Wagner D: **Towards evaluating the robustness of neural networks**. In *Proceedings - IEEE symposium on security and privacy*; 2017:39–57. https://doi.org/10.1109/SP.2017.49.

This paper critiques defensive distillation, a method intended to enhance neural network robustness against adversarial examples. The study introduces three new attack algorithms that effectively target both distilled and undistilled networks, challenging the perceived effectiveness of defensive distillation. Additionally, the paper suggests using high-confidence adversarial examples for testing and breaking defensive distillation, aiming to provide a benchmark for evaluating future defenses.

8. * Finlayson SG, Chung HW, Kohane IS, Beam AL: ***Adversarial attacks against medical deep learning systems***. http://arxiv.org/abs/1804.05296; 2018.

This paper highlights the emerging issue of adversarial attacks in machine learning systems used in healthcare. It discusses how these attacks can drastically alter system outputs with minor input changes, posing significant risks in a high-stakes environment. The paper calls for collaborative efforts among medical, technical, legal, and ethical experts to address these vulnerabilities and ensure the secure and effective use of machine learning in healthcare.

9. * Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R: **Intriguing properties of neural networks**. In *2nd international conference on learning representations, ICLR 2014 - conference track proceedings*; 2014.

This paper explores two critical properties of deep neural networks. Firstly, it finds that high-level units in these networks do not uniquely contribute to semantic information; rather, it is the combined space of these units that holds the information. Secondly, it reveals that neural networks often learn input-output mappings that are discontinuous, making them vulnerable to adversarial perturbations. These perturbations can cause misclassification and are transferable across dicerent networks trained on varied datasets.

10. Yuan X, He P, Zhu Q, Li X: **Adversarial examples: attacks and defenses for deep learning**. *IEEE Transact Neural Networks Learn Syst* 2019, **30**. https://doi.org/10.1109/TNNLS.2018.2886017.
*
This paper reviews recent developments in adversarial examples, which are subtle in-put perturbations that can deceive deep neural networks despite being imperceptible to humans. It summarizes various methods for generating these adversarial examples, proposes a taxonomy for categorizing these methods, and discusses their applications. Additionally, the paper explores countermeasures to defend against adversarial attacks and addresses the associated challenges and potential solutions.

11. Kurakin A, Goodfellow IJ, Bengio S: **Adversarial machine learning at scale**. In *5th international conference on learning representations, ICLR 2017 - conference track proceedings*; 2017.
*
This research explores the application of adversarial training to large-scale models and datasets, specifically ImageNet. It investigates how adversarial training can be scaled effectively, addresses the robustness of models against single-step versus multi-step attacks, and resolves issues related to label leakage in adversarial training. The study provides valuable insights into enhancing model resilience to adversarial examples and improving the effectiveness of training techniques in large-scale scenarios.

12. Mkadry A, Makelov A, Schmidt L, Tsipras D, Vladu A: **Towards deep learning models resistant to adversarial attacks**. *Stat* 2017, **1050**.
*
This paper explores neural network vulnerability to adversarial examples and proposes a robust optimization approach to enhance adversarial robustness. It offers a unifying framework for understanding and improving neural network security, providing methods for training and attacking that offer concrete security guarantees. The study highlights the importance of robustness against first-order adversaries as a key step toward creating resilient deep learning models.

13. Croce F, Hein M: **Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks**. In *International conference on machine learning*; 2020:2206–2216.
*
This paper addresses issues in evaluating adversarial defenses by introducing improved versions of the PGD attack to overcome common pitfalls such as suboptimal step sizes and objective function problems. The authors propose an ensemble of attacks that combines their novel methods with existing ones to provide a comprehensive and robust evaluation of adversarial defenses. Applying this ensemble to over 50 models from recent top publications, they frequently achieve lower robust test accuracy than reported, highlighting the potential overestimation of defense effectiveness in previous studies.

14. Fan Y, Wu B, Li T, Zhang Y, Li M, Li Z, Yang Y: **Sparse Adversarial Attack via Perturbation Factorization**. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII*. Berlin, Heidelberg: Springer-Verlag; 2020:35–50. https://doi.org/10.1007/978-3-030-58542-6_3.
*
This paper introduces a novel sparse adversarial attack approach that optimizes perturbations at selected pixel positions to mislead a deep neural network. Unlike previous methods that manually select perturbation positions, the proposed method factors perturbations into magnitudes and binary selection factors, modeling the problem as a MIP problem. This approach allows joint optimization of selection and magnitude, with constraints to control sparsity and enhance visual imperceptibility. An efficient algorithm reformulates the MIP as a continuous optimization problem. Experimental results show that this method outperforms existing sparse attack techniques.

15. Dziugaite GK, Ghahramani Z, Roy DM: ***A study of the effect of JPG compression on adversarial images*** (Isba), http://arxiv.org/abs/1608.00853; 2016.
*
This study examines how JPG compression affects the classification of adversarial images-inputs altered to mislead neural networks. The research finds that JPG compression can significantly reduce the impact of small-magnitude adversarial perturbations on classification accuracy, though it is less effective for larger perturbations. This highlights the complex interplay between image compression and adversarial robustness, revealing that JPG compression can sometimes mitigate but not fully counteract the adversarial effects.

16. Zantedeschi V, Nicolae MI, Rawat A: **Efficient defenses against adversarial attacks**. In *AISec 2017 - proceedings of the 10th ACM workshop on artificial intelligence and security, co-located with CCS 2017*; 2017:39–49. https://doi.org/10.1145/3128572.3140449.
*
This research investigates e cient defenses against adversarial attacks, contributing to the development of strategies to enhance the robustness of machine learning models. The study is presented in the context of artificial intelligence and security, providing insights into practical methods to mitigate the impact of adversarial challenges on machine learning systems.

17. Guo C, Rana M, Cissé M, Van Der Maaten L: *Countering adversarial images using input transformations. 6th international conference on learning representations, ICLR 2018 - conference track proceedings*. 2018.
*
This work explores the use of input transformations to counter adversarial images. The authors contribute to the field of adversarial robustness by investigating techniques that involve modifying input data to enhance the resilience of machine learning models. The study offers valuable insights into practical methods for mitigating the impact of adversarial attacks.

18. Welling M: *Meta-learning for medical image classification. (1549)*. 2018:7–9.
*
In this work, Max Welling explores the application of meta-learning techniques for medical image classification. The study contributes to the field of machine learning in healthcare by investigating methods that leverage meta-learning to enhance the e ciency and adaptability of models for medical image analysis, potentially improving diagnostic accuracy and treatment planning.

19. Xu W, Evans D, Qi Y: ***Feature squeezing: detecting adversarial examples in deep neural networks***. February, https://doi.org/10.14722/ndss.2018.23198; 2018.
*
This research introduces Feature Squeezing as a method for detecting adversarial examples in deep neural networks. The authors contribute to the development of techniques aimed at identifying and mitigating adversarial attacks, enhancing the security and reliability of deep learning models. The study provides insights into practical approaches for detecting adversarial examples in the context of deep neural networks.

20. Buckman J, Roy A, Raffel C, Goodfellow I: **Thermometer encoding: one hot way to resist adversarial examples**. In *6th international conference on learning representations, ICLR 2018 - conference track proceedings*; 2018:1–22. 2016.
*
In this work, the authors propose Thermometer Encoding as a method to resist adversarial examples. The study contributes to the field of adversarial robustness, providing insights into a technique that aims to enhance the resilience of machine learning models against adversarial attacks.

21. Miyato T, Maeda SI, Koyama M, Ishii S: **Virtual adversarial training: a regularization method for supervised and semi-supervised learning**. *IEEE Trans Pattern Anal Mach Intell* 2019, **41**. https://doi.org/10.1109/TPAMI.2018.2858821.
*
This paper introduces Virtual Adversarial Training, a regularization method for both supervised and semi-supervised learning. The authors contribute to the field of machine learning by proposing a technique that leverages adversarial training for model regularization, demonstrating its applicability in both fully and partially labeled datasets.

22. Nicolae M, Sinn M, Tran MN, Buesser B, Rawat A, Wistuba M, Zantedeschi V, Baracaldo N, Chen B, Ludwig H, Molloy IM, Edwards B: **Adversarial robustness toolbox v1.0.0**. *ArXiv* 2018. https://arxiv.org/abs/1807.01069.
*
This work represents the release of the Adversarial Robustness Toolbox version 1.0.0, a comprehensive resource for researchers and practitioners working on adversarial machine learning. The toolbox is designed to provide tools and functionalities for evaluating and enhancing the robustness of machine learning models against adversarial attacks. The authors contribute to the development of resources aimed at addressing the challenges associated with adversarial examples in machine learning.

23. Goodman D, Xin H, Yang W, Yuesheng W, Junfeng X, Huan Z: **Advbox: a toolbox to generate adversarial examples that fool neural networks**. *ArXiv* 2020. https://arxiv.org/abs/2001.05574.
*
This paper introduces Advbox, a comprehensive toolbox for generating adversarial examples to test the robustness of neural networks across multiple frameworks, including PaddlePaddle, PyTorch, and TensorFlow. Advbox supports various attack scenarios, such as black box attacks and specific applications like Face Recognition Attack and DeepFake Face Detection. The toolbox, licensed under Apache 2.0 and compatible with Python 3, provides a versatile platform for evaluating and benchmarking machine learning model vulnerabilities.

24. Papernot N, McDaniel P, Wu X, Jha S, Swami A: **Distillation as a defense to adversarial perturbations against deep neural**
*

networks. In *Proceedings - 2016 IEEE symposium on security and privacy, SP 2016*; 2016. https://doi.org/10.1109/SP.2016.41.
This paper explores the use of distillation as a defense mechanism against adversarial perturbations targeting deep neural networks. The authors contribute to the field of adversarial machine learning by proposing distillation, a technique that involves training a smaller model to mimic the behavior of a larger model, as a defense strategy.

25. Moosavi-Dezfooli SM, Fawzi A, Frossard P: **DeepFool: a simple and accurate method to fool deep neural networks**. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition, 2016-Decem*; 2016:2574–2582. https://doi.org/10.1109/CVPR.2016.282.
*
This work introduces DeepFool, a method designed to generate adversarial examples with simplicity and high accuracy, aiming to deceive deep neural networks. The authors contribute to the understanding and exploration of adversarial attacks on deep learning models, providing a valuable tool for evaluating the robustness of neural networks in computer vision tasks.

26. Ling X, Ji S, Zou J, Wang J, Wu C, Li B, Wang T: **Deepsec: a uniform platform for security analysis of deep learning model**. In *2019 IEEE symposium on security and privacy (SP)*; 2019:673–690.
*
This paper introduces DEEPSEC, a comprehensive platform designed for evaluating adversarial attacks and defenses in deep learning models. DEEPSEC integrates 16 stateof-the-art attacks and 13 defenses with various utility metrics, enabling researchers to assess model vulnerability and defense effectiveness systematically. The platform addresses critical questions in adversarial learning and reveals insights such as the limitations of universal defenses and the impact of defense ensembles. DEEPSEC serves as a valuable tool for advancing research in adversarial machine learning.

27. Simonyan K, Zisserman A: ***Very deep convolutional networks for large-scale image recognition***. https://arxiv.org/abs/1409.1556; 2015.
*
This work explores the impact of convolutional network depth on accuracy in largescale image recognition, using very small (3x3) convolution filters. The study demonstrates that increasing network depth to 1619 layers significantly improves performance. These findings contributed to securing top positions in the ImageNet Challenge 2014. The paper also highlights the generalizability of these deep visual representations across other datasets and makes the top-performing ConvNet models publicly available for further research.