ADVERSARIALLY ROBUST DENSENET FOR CHEST X-RAY CLASSIFICATION WITH PGD ATTACK AND PIXEL DEFLECTION TRANSFORM

JEYANTH SIDDHARTH SEEKA CHITTI, URVI HAVAL, VENKATESWARA RAO KODURI

ABSTRACT. **Adversarial attacks pose a significant threat to deep learning models, especially in high-stakes domains such as medical imaging. A convolutional deep learning model is not considered robust in the current landscape unless it demonstrates sufficient resilience to adversarial attacks. Implementing adversarial defense methods is particularly critical in medical imaging, where reliability and accuracy are paramount. Although achieving a system that is 100% robust remains a challenge, our objective is to develop a well-rounded model that maintains classification accuracy and mitigates the risk of disease misdiagnosis, even in the presence of adversarial manipulations. In this project, we have utilized the DenseNet121 model trained on the CheXpert dataset, which involves multi-label X-ray classifications. It focuses on generating adversarial examples using Projected Gradient Descent (PGD) and mitigating their impact using a combination of Robust Class Activation Maps (RCAM) and Pixel Deflection Transform (PDT).**

1. INTRODUCTION

Deep learning has revolutionized the field of medical imaging, enabling automated and highly accurate diagnosis of diseases through modalities such as chest X-rays, CT scans, and MRIs [1, 8, 13]. Several deep learning models, including DenseNet, ResNet, EfficientNet, and Vision Transformers, have demonstrated strong performance in multi-modal classification of medical images, positioning them as valuable tools for assisting clinicians in decision-making. However, their vulnerability to adversarial attacks—intentional input manipulations designed to deceive models—raises significant concerns, particularly in high-stakes applications such as healthcare.

Adversarial attacks can introduce small, imperceptible perturbations to input images, causing models to misclassify them with high confidence. This weakness, if exploited, can compromise the reliability and safety of AI-based medical diagnostic systems. To counter this, researchers have proposed various defense mechanisms, including adversarial training and preprocessing methods to improve model robustness. However, achieving a balance between robustness and accuracy remains a challenge.

In this project, we explore the robustness of DenseNet121 [4], a popular convolutional neural network architecture, on a subset of the CheXpert dataset—a large-scale chest X-ray dataset designed for multi-label classification of thoracic diseases [6]. We employ **Projected Gradient Descent (PGD)** [7], a widely used adversarial attack method, to evaluate the susceptibility of the trained model to adversarial inputs. To defend against this attack, we incorporate two key strategies: **Robust Class Activation Maps (RCAM)** [8, 13] and **Pixel Deflection Transform (PDT)** [8]. RCAM identifies critical regions of the image for classification, allowing us to selectively apply defenses, while PDT perturbs non-critical pixels to mitigate adversarial noise.

The pipeline involves the following steps:

(1) Training **DenseNet121** on the **CheXpert** dataset and evaluating its performance on clean data.
(2) Generating adversarial examples using **Projected Gradient Descent (PGD)** to test the model's robustness.
(3) Defending against adversarial attacks using **RCAM-guided PDT** and adversarial training.

Performance is evaluated using metrics such as AUC-ROC, clean accuracy, adversarial accuracy, and the overall robustness of the model. By systematically integrating attack and defense mechanisms into the training and evaluation pipeline, this work aims to explore the resilience of deep learning models in the high-stakes domain of medical imaging.

2. BACKGROUND

This section introduces the foundational concepts underlying our work, including the DenseNet-121 architecture, Projected Gradient Descent (PGD) as an adversarial attack method, and Pixel Deflection Transform (PDT) as a defense mechanism.

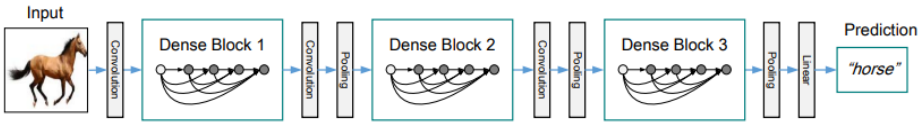

FIGURE 1. A deep DenseNet with three dense blocks from the original implementation

DenseNet-121 Architecture:
DenseNet-121 is a convolutional neural network architecture part of the DenseNet family, known for its efficient utilization of parameters and computational resources. DenseNet introduces dense connectivity, unlike traditional convolutional networks where layers are connected sequentially. In this architecture, each layer is connected to every subsequent layer, ensuring that feature maps learned by earlier layers are directly accessible to later layers. This connectivity mitiagtes the vanishing gradient problem, allowing deeper networks with fewer parameters. The '121' in DenseNet-121 refers to the number of layers in the network. This architecture has been widely adopted for medical image analysis tasks due to its ability to capture intricate patterns in high-resolution images, making it a suitable choice for our study. [4]

Projected Gradient Descent (PGD):
Projected Gradient Descent (PGD) is a widely studied adversarial attack method. It is an iterative, first-order optimization algorithm that generates adversarial examples by perturbing input images within a predefined norm constraint. PGD aims to maximize the model's loss function with respect to the input image, while ensuring that the perturbation remains imperceptible to human observers. This is achieved by iteratively updating the adversarial image using the gradient of the loss function and projecting the perturbed image back into the L∞ ball defined by a perturbation bound ϵ. Mathematically, the update rule can be expressed as:

$$\mathbf{adv\_images}_{t+1} = \text{clip}_{\mathbf{images},\epsilon}\left\{\mathbf{adv\_images}_t + \alpha \cdot \text{sign}\left(\nabla_{\mathbf{adv\_images}}\mathcal{L}(\mathbf{outputs}, \mathbf{labels})\right)\right\}$$

Where:
- $\mathbf{adv\_images}_t$: Adversarial image at iteration t.
- α: Step size for updating the adversarial example.
- $\mathcal{L}$: Loss function (e.g., Cross-Entropy Loss).
- $\mathbf{outputs}$: Model predictions for $\mathbf{adv\_images}_t$.
- $\mathbf{labels}$: True label for the input image.
- $\nabla_{\mathbf{adv\_images}}\mathcal{L}$: Gradient of the loss function with respect to the adversarial image.
- $\text{clip}_{\mathbf{images},\epsilon}$: Clipping function that ensures the perturbation remains within the L_∞-norm constraint ϵ and valid pixel intensity bounds $[0, 1]$.

Pixel Deflection Transform (PDT):
Pixel Deflection Transform (PDT) is a simple yet effective defense mechanism introduced to mitigate the effects of adversarial perturbations in images. It operates by disrupting the structured noise introduced by adversarial attacks, effectively reducing the model's susceptibility to adversarial examples. PDT achieves this by randomly altering pixel values in the image, deflecting them to new locations within a specified neighborhood. Mathematically, for a given image $\mathbf{I}$ of dimensions $H \times W \times C$ (height, width, and channels), a pixel at position (x, y) in the deflected image $\mathbf{I}'$ is updated as:

$$\mathbf{I}'(x, y, c) = \mathbf{I}(x + \Delta x, y + \Delta y, c),$$

where $\Delta x, \Delta y$ are random offsets sampled from a uniform distribution within the range:

$$\Delta x, \Delta y \sim \text{Uniform}(-\text{window}, \text{window}).$$

The selected neighboring pixel $(x + \Delta x, y + \Delta y)$ must lie within the image bounds, ensuring:

$$0 \leq x + \Delta x < H \quad \text{and} \quad 0 \leq y + \Delta y < W.$$

3. RELATED WORK

The study of adversarial robustness in deep learning has garnered significant attention, especially in safety-critical domains such as medical imaging. [10, 9] Adversarial attacks, first highlighted by [11] and further modified by

Goodfellow et al. (2014) [3] through the introduction of Fast Gradient Sign Method (FGSM), demonstrate how small, imperceptible perturbations can drastically degrade model performance and lead to misclassifications. Among these attacks, Projected Gradient Descent (PGD), as proposed by Madry et al. (2017) [7], is considered one of the most effective first-order adversarial methods, iteratively optimizing perturbations within an $L\infty$ norm constraint. This work demonstrated that adversarial training—where a model is explicitly trained on adversarial examples—can significantly improve a model's robustness to adversarial attacks. However, this method is computationally expensive due to the iterative nature of generating adversarial examples during training. DenseNet-121, a densely connected convolutional neural network introduced in [5], has been widely adopted for medical image analysis, including applications in chest X-ray interpretation. Despite its performance advantages, DenseNet-121 remains susceptible to adversarial perturbations, necessitating robust defenses.

To address these challenges, various adversarial defenses have been proposed. One such approach is the Pixel Deflection Transform (PDT), introduced by Prakash et al. (2018) in [8], which aims to disrupt adversarial noise by randomly redistributing pixel values within localized neighbourhoods. PDT is a lightweight, post-hoc defense that does not require re-training the model, making it computationally efficient compared to adversarial training. Other defense mechanisms, such as randomized smoothing [2] and feature squeezing [12], have also been explored, but each comes with trade-offs in computational overhead and applicability to large-scale datasets like CheXpert. Medical imaging datasets pose unique challenges due to their high dimensionality and the clinical importance of preserving image integrity. The CheXpert dataset, proposed in [6], is one of the largest labeled medical imaging datasets, commonly used as a benchmark for chest X-ray classification tasks.

In this context, our work seeks to build on these advancements by focusing on the application of PGD attacks and PDT defenses within the medical imaging domain. While adversarial training remains the gold standard for robustness, we explore the potential of PDT as a practical and efficient alternative for defending DenseNet-121 against adversarial perturbations. By leveraging a subset of the CheXpert dataset and focusing on evaluation rather than re-training, our study provides insights into the trade-offs between robustness, computational feasibility, and model accuracy in the context of adversarial attacks and defenses in medical imaging.

4. Approach

Our approach began with an exploration of a two-step methodology: first, training a DenseNet-121 model on clean images from the CheXpert dataset and subsequently leveraging adversarial training to enhance the model's robustness. However, this initial approach posed significant computational challenges. The CheXpert dataset consists of approximately 222,000 chest X-ray images, which together amount to over 11 GB of data. Training DenseNet-121 on this large dataset, including the additional step of adversarial training with PGD-generated images, was estimated to require over 55 hours of computational time on AWS GPU instances. Given these resource-intensive requirements, it became apparent that this approach was not feasible within our constraints. This necessitated a shift toward a more computationally efficient strategy while retaining the core objectives of understanding the impact of adversarial attacks and evaluating the efficacy of PDT as a defense mechanism.

In the revised methodology, the training dataset was randomly downsampled to a subset of 15,000 images, ensuring that it remained representative of the broader CheXpert dataset in terms of class balance and diversity. Instead of re-training the DenseNet-121 model on adversarial examples, we decided to leverage the model trained solely on clean images as the baseline. This allowed us to focus our efforts on understanding the impact of adversarial perturbations and the subsequent mitigation strategies without the additional complexity of adversarial training.

The revised pipeline involves three key steps. First, we established baseline performance by evaluating our hardcoded DenseNet-121 model on clean images from the reduced dataset and recording the model performance. Second, we applied PGD (Projected Gradient Descent) to the clean images to generate adversarial examples. This step enabled us to evaluate the model's vulnerability to adversarial attacks. Third, we implemented the Pixel Deflection Transform as a defense mechanism against the PGD attack. PDT was applied to both clean and adversarial images, and the DenseNet-121 model's performance on these transformed images was evaluated.

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

FIGURE 2. CheXpert Dataset

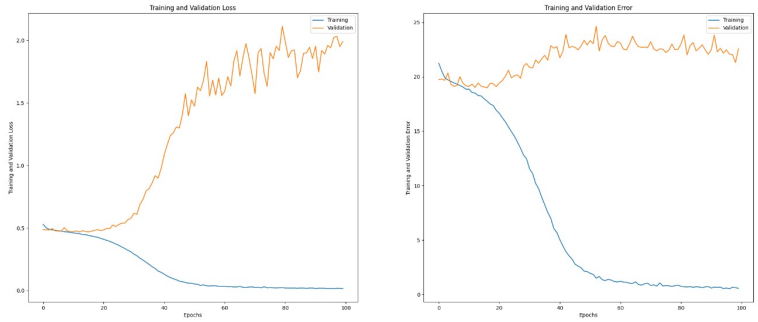

FIGURE 3. Training and validation performance for 100 epochs

5. EXPERIMENTAL RESULTS

About the CheXpert Dataset: The CheXpert dataset used in our study includes annotations for 14 lung conditions and characteristics. Unlike binary classification datasets, where labels are typically ”yes” or ”no,” the labels in CheXpert can take three values: positive, negative, or uncertain.

During preprocessing, we encountered missing values in the dataset’s CSV file for certain lung conditions and characteristics. These missing entries were treated as ”uncertain” and assigned a value of 2. Specifically, we standardized the labels as follows: positive conditions were labeled as 1, negative conditions as 0, and uncertain or missing values as 2. While this preprocessing step ensured consistency across the dataset, it may have introduced noise into the model’s training process.

When training the DenseNet-121 model on a reduced dataset of 15,000 images with a 70:30 train-test split and 100 epochs, we encountered significant overfitting. This overfitting can be attributed to the high complexity of the DenseNet-121 architecture relative to the limited size of the training dataset. DenseNet-121, with its densely connected convolutional layers, is designed to capture intricate patterns and relationships in large datasets. However, when trained on a smaller dataset, the model tends to memorize the training data rather than generalize effectively to unseen data. Despite the constraints of having to train on a smaller dataset, the reduction in data size was necessary due to computational limitations.

To address the overfitting issue observed during training, we experimented with various strategies aimed at improving the model’s generalization. One of the first steps we took was to increase the size of the validation set, effectively adjusting the split ratio to provide more data for validation. However, this change did not lead to any significant improvement, as the overfitting persisted. We then experimented with a learning rate scheduler to dynamically adjust the learning rate during training, hoping to stabilize the optimization process and reduce overfitting, Figure 5 shows the

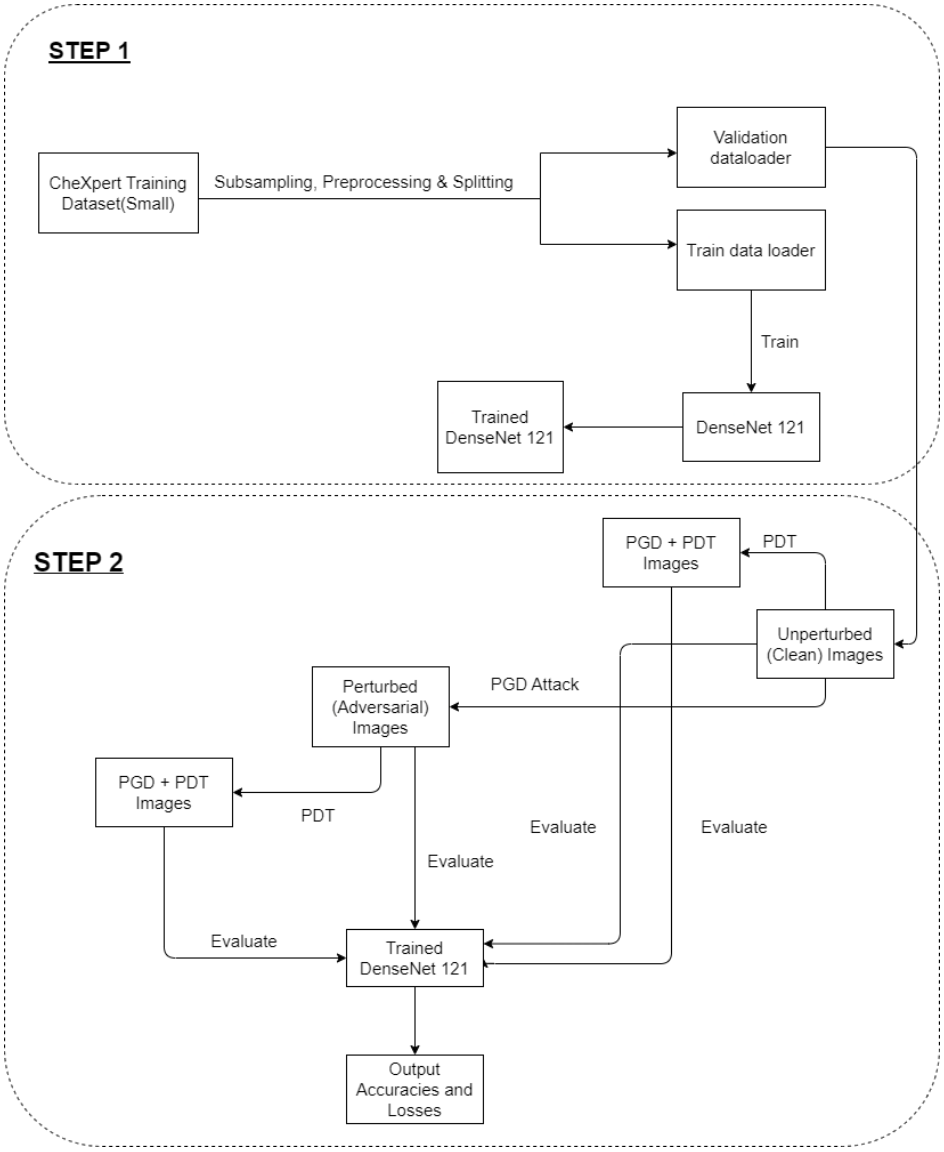

FIGURE 4. Execution Flowchart

model performance after the learning rate scheduler. We can see here that even the Learning Rate Scheduler does not reduce the overfitting. Figure 6 shows the training and validation performance after the final model was trained.

6. DISCUSSION

According to the obtained plots shown in Figure 7 and Figure 8, on clean images, the model achieves a validation loss of 0.422, indicating its ability to effectively generalize on unperturbed data. However, when evaluated on adversarial images, the validation loss rises significantly to 0.802, highlighting the model's vulnerability to adversarial perturbations. Introducing the Pixel Deflection Transform (PDT) as a defense reduces the validation loss on adversarial images to 0.785, demonstrating that PDT mitigates the impact of adversarial attacks, though the improvement is modest. Interestingly, when PDT is applied to clean images, the validation loss remains nearly unchanged at 0.424 compared to the original clean loss, suggesting that PDT does not adversely affect the model's performance on clean data.

In future work, we aim to address the limitations imposed by the constraints in our current study. One significant improvement we plan to make is training the model on the entire CheXpert dataset instead of the reduced subset of 15,000 images. By utilizing the full dataset, we believe the DenseNet-121 model will be able to better capture the complexities of the data, which could improve its generalization and reduce overfitting. Additionally, we intend to focus

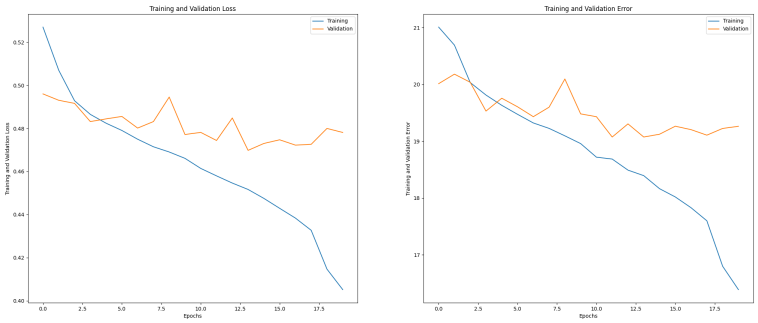

FIGURE 5. Training and validation performance for 20 epochs

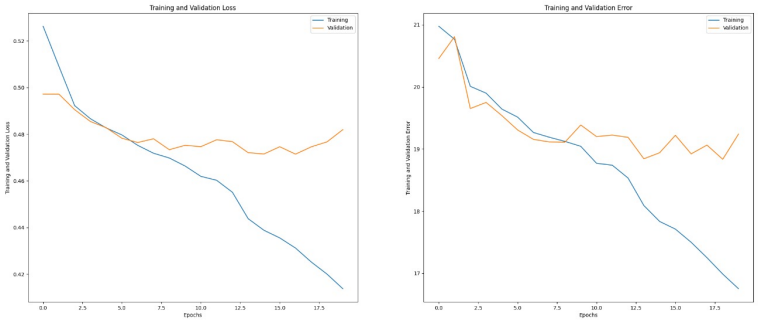

FIGURE 6. Train and Validation Performance after model training

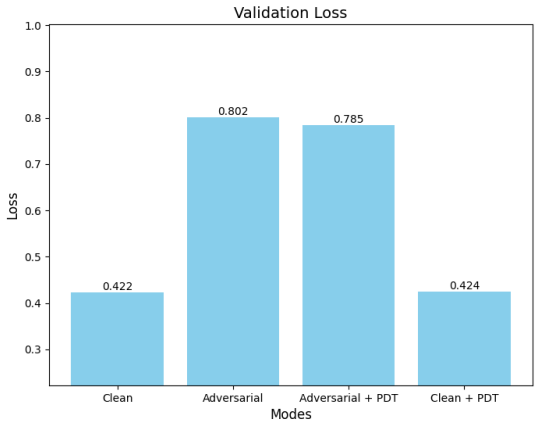

FIGURE 7. Validation Loss Bar graph

on fine-tuning the parameters of the Pixel Deflection Transform (PDT) defense. By experimenting with the number of deflections and the size of the deflection window, we hope to further enhance the model's robustness against adversarial attacks and achieve better performance.

REFERENCES

[1] M.A. Abdou. Literature review: efficient deep neural networks techniques for medical image analysis. *Neural Computing and Applications*, 34(8):5791–5812, 2022.

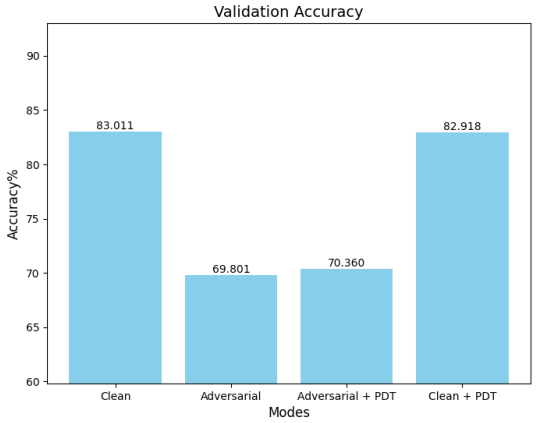

FIGURE 8. Validation Accuracy Bar Graph

[2] Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing, 2019.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv e-prints*, page arXiv:1412.6572, December 2014.

[4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

[6] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, Daniel A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[8] Aaditya Prakash, Nick Moran, Solomon Garber, Ruiyu Hu, and Hannaneh Hajishirzi. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8571–8580, 2018.

[9] David Rodriguez, Tapsya Nayak, Yidong Chen, Ram Krishnan, and Yufei Huang. On the role of deep learning model complexity in adversarial robustness for medical images. *BMC Medical Informatics and Decision Making*, 22(2):160, 2022.

[10] Wenjie Ruan, Xinping Yi, and Xiaowei Huang. Adversarial robustness of deep learning: Theory, algorithms, and applications. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4866–4869, New York, NY, USA, 2021. Association for Computing Machinery.

[11] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[12] Guiyong Xu, Yang Xu, Sicong Zhang, and Xiaoyao Xie. Sfrnet: Feature extraction-fusion steganalysis network based on squeeze-and-excitation block and repvgg block. *Security and Communication Networks*, 2021(1):3676720, 2021.

[13] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.