NAMSCON 2023

RAMAIAH Medical College | A Constituent College, Ramaiah University of Applied Sciences

Enhancing Diagnostic Accuracy by Remediation of Adversarial Attacks on Deep Learning-Based Neuroimaging Systems

Mayur Mankar[1,3*], Parth Joshi[2] and Prasun K. Roy[1,4#]

[1]Neuroimaging Laboratory, Dept. of Life Sciences, Shiv Nadar University (UGC Institution of Eminence), Delhi NCR, 201314

[2]Dept. of Electrical Engineering, Shiv Nadar University (UGC Institution of Eminence), Delhi NCR, 201314

[3]Dept. of Data Science and Engineering, Indian Institute of Science Education and Research, Bhopal, 462066

[4]SNU-Dassault Centre of Excellence on Research & Innovation, Shiv Nadar University, Greater Noida, Delhi NCR, 201314

Corresponding author: prasun.roy@snu.edu.in

Introduction

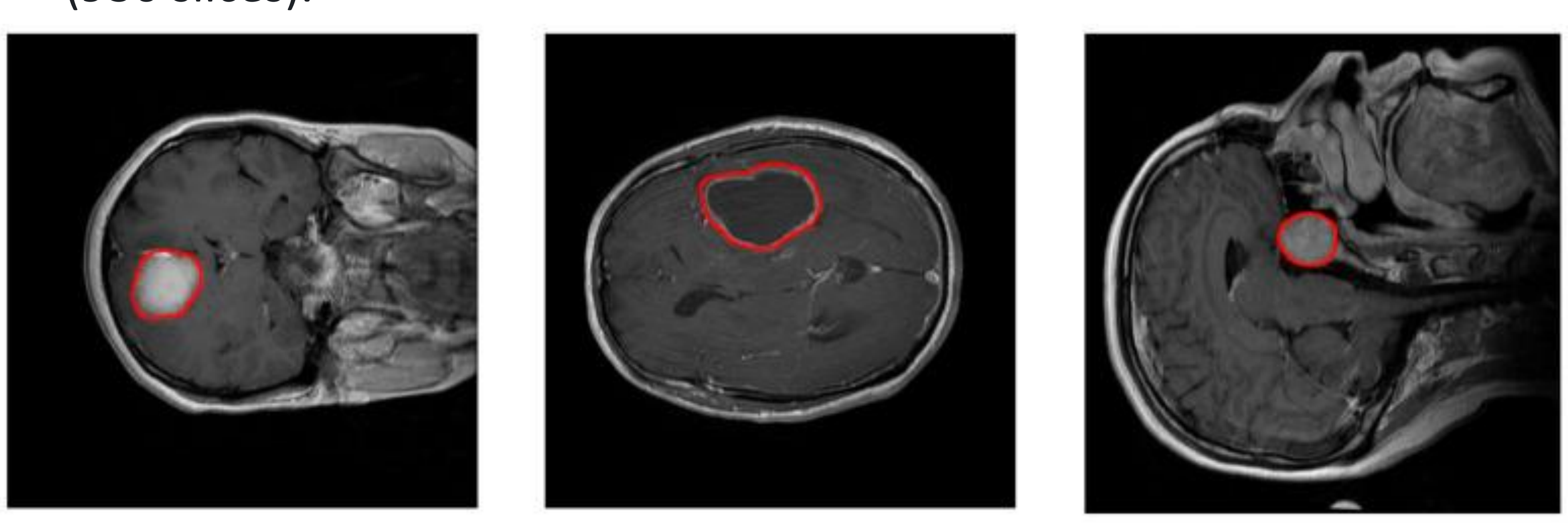
- Deep Neural Networks have become well-utilized for medical image analysis tasks like brain lesions, atrophy or tumor detection, diagnosis, and grading.
- However, recent studies demonstrate that carefully-engineered adversarial attacks can compromise medical deep learning systems with small imperceptible perturbations.
- This raises security concerns about the deployment of deep learning-based systems in clinical settings.

Objective

- Our study investigates the robustness of deep learning-based MRI diagnostic systems using adversarial images and looks into an iterative adversarial training approach to defence against these attacks.

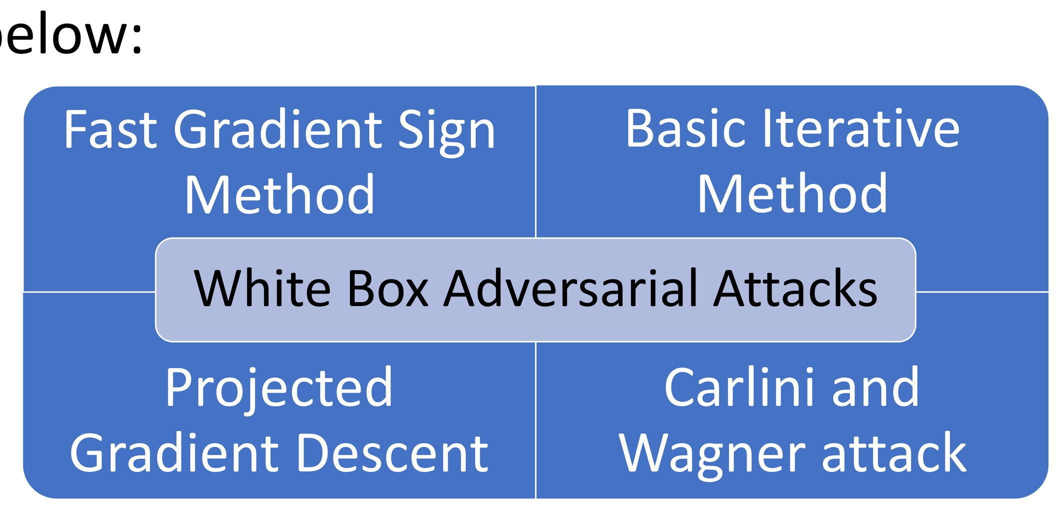
Materials & Methods

- We used the brain tumor MRI platform of 3064 T1-weighted contrast-enhanced images from 233 patients with three kinds of brain tumor: meningioma (708 slices), glioma (1426 slices), and pituitary tumor (930 slices).



Examples of MRI images of the T1-CE MRI image dataset. Left: coronal view of a meningioma tumor. Center: Axial view of a glioma tumor. Right: sagittal view of a pituitary tumor. Tumor borders have been highlighted in red.

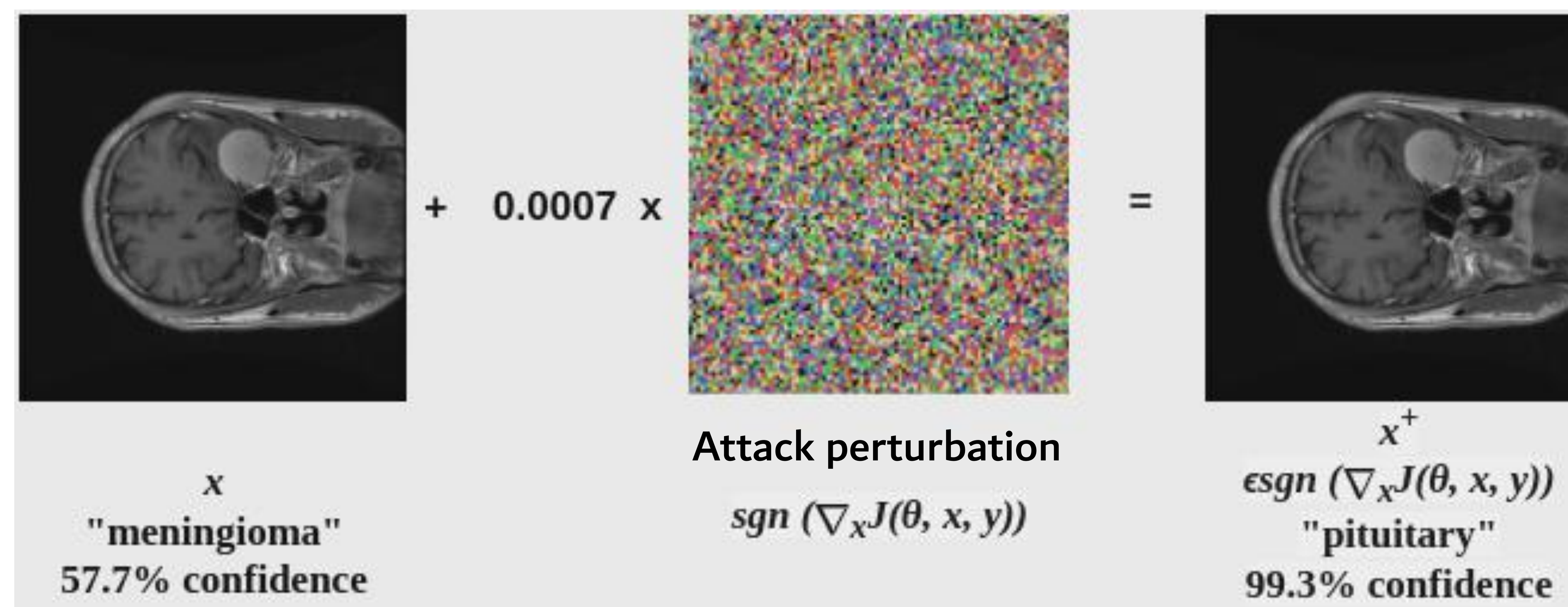
- We examined the impact of adversarial attacks on the **classification accuracy** of an ImageNet (natural images) pre-trained **ResNet-50** model as the base network.
- The model is trained to classify brain tumors in T1-weighted contrast-enhanced MRI images into **meningioma**, **glioma**, and **pituitary** tumor.
- Small perturbation-based white-box adversarial attack is applied on testing MRI images to check the accuracy of trained model on perturbed MRI images.
- We explored four different white-box adversarial attacks as shown in the schema below:



Schema 1: White-box adversarial attacks

- Then, to remedy this attack, we now pursued the utility of an iterative adversarial training approach to improve the robustness of this system against adversarial images in the context of MRI images.
- We applied the Projected Gradient Descent (PGD) attack method to the training dataset so that convolutional neural network can learn to ignore the noise patterns.

Procedure of Adversarial Attacks



Results

- Initial diagnostic accuracy of the deep neural network to differentially identify glioma, meningioma, and pituitary tumor is **99.86%.**
- Very small pixel-level perturbation $\varepsilon = 0.04$ resulted in sharp decrease in accuracy (**FGSM 29.98%, BIM 0.21%, PGD 0.43%, CW 0.21%**).
- Strong attack methods like BIM, PGD, and CW do damage by minimal perturbations, highlighting that targeting medical images is notably less challenging compared to natural image datasets like CIFAR-10 and ImageNet.
- In the case of FGSM attack, a considerably larger perturbation is typically needed to succeed.
- Adversarial training slightly lowered the classification accuracy at baseline (on non-perturbed images) from **99.86%** to **99.24%**.
- Adversarial training improved the robustness of deep neural network; accuracy increased from **29.98%** to **96.14%** in case of PGD attack.

Conclusion

- Deep learning-based diagnostic systems has high prediction accuracy but are vulnerable to malicious adversarial attacks.
- These diagnostic systems naively trained on medical images exhibited dramatic instability to small pixel-level changes resulting in a huge decrease in accuracy.
- Adversarial training techniques improved the stability and robustness of deep neural network to such pixel-level changes.
- Despite adversarial training, deep neural network did not reach baseline accuracy, suggesting adversarial training as only a partial solution to improve model robustness.

References

[1] He K. et al, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, 2016.

[2] Cheng J. et al, "Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition". PLoS ONE 10(10): e0140381, 2015.

[3] Goodfellow I et al. "Explaining and Harnessing Adversarial Examples." In 3rd International Conference on Learning Representations, ICLR 2015, 2015, Ed. Y Bengio, Y LeCun.

[4] Kurakin A et al. "Adversarial Examples in the Physical World." CoRR abs/1607.02533, 2016.

[5] Madry A et al, Towards Deep Learning Models Resistant to Adversarial Attacks. Openreview.net. 2018 .

[6] N. Carlini et al, "Towards Evaluating the Robustness of Neural Networks," 2017 IEEE Symposium on Security and Privacy (SP), 2017, pp. 39-57.

[7] D A et al. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." In *9th International Conference on Learning Representations, ICLR 2021,* OpenReview.net.

Acknowledgment

The authors greatly appreciate the support from C3i Hub Foundation (sponsored by DST: Dept. of Science & Technology, Govt. of India).

#827