scientific reports

Check for updates

OPEN

A multi-layered defense against adversarial attacks in brain tumor classification using ensemble adversarial training and feature squeezing

Ahmeed Yinusa[1] & Misa Faezipour[2]✉

Deep learning, particularly convolutional neural networks (CNNs), has proven valuable for brain tumor classification, aiding diagnostic and therapeutic decisions in medical imaging. Despite their accuracy, these models are vulnerable to adversarial attacks, compromising their reliability in clinical settings. In this research, we utilized a VGG16-based CNN model to classify brain tumors, achieving 96% accuracy on clean magnetic resonance imaging (MRI) data. To assess robustness, we exposed the model to Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks, which reduced accuracy to 32% and 13%, respectively. We then applied a multi-layered defense strategy, including adversarial training with FGSM and PGD examples and feature squeezing techniques such as bit-depth reduction and Gaussian blurring. This approach improved model resilience, achieving 54% accuracy on FGSM and 47% on PGD adversarial examples. Our results highlight the importance of proactive defense strategies for maintaining the reliability of AI in medical imaging under adversarial conditions.

Keywords Brain tumor, Magnetic resonance imaging (MRI), Artificial intelligence (AI), Adversarial attacks, Adversarial training, Feature squeezing

In recent years, advancements in medical imaging, especially Magnetic Resonance Imaging (MRI), have revolutionized brain tumor diagnosis and evaluation[1]. MRI stands out in brain tumor imaging for its high-resolution capabilities and non-use of ionizing radiation, effectively differentiating tissue types[2]. MRI provides vital information about tumor location, size, and shape, aiding clinicians in planning surgeries and other treatments. However, analyzing MRI scans for brain tumors can be challenging due to the subtle differences in texture, intensity, and shape that differentiate tumor types[3]. Manual analysis of MRI images is time-intensive and prone to human error, which has driven researchers toward automated, AI-driven classification approaches[4].

Artificial intelligence (AI), especially deep learning, has transformed many domains of medical imaging analysis, including skin cancer detection[5] and diabetic retinopathy classification[6]. Recent research has further demonstrated the effectiveness of deep learning and radiomics in other critical diagnostic areas, such as Alzheimer's disease classification using multi-modal neuroimaging[7] and breast cancer diagnosis through multiparametric mammography[8]. Convolutional Neural Networks (CNNs) are a type of deep learning architecture that has achieved remarkable success in medical image classification due to their ability to learn complex hierarchical patterns from large datasets[9,10].

CNNs are particularly effective in tasks like brain tumor classification, where nuanced visual differences may indicate significant clinical outcomes. When trained on comprehensive datasets, CNNs can recognize patterns specific to tumor types and grades, sometimes matching or even exceeding human diagnostic accuracy[11].

Several recent works have proposed advanced deep learning-based approaches for MRI-based brain tumor detection. Ullah et al. introduced a transfer learning approach to effectively detect and identify brain tumors, improving classification accuracy[12]. Additionally, a robust end-to-end deep learning model was developed to ensure reliable and efficient brain tumor diagnosis using MR images[13]. Furthermore, enhancing model transparency, Ullah et al. proposed the DeepEBTDNet model combined with LIME, providing explainability

[1]Computational and Data Science Program, Middle Tennessee State University, 1301 East Main Street, Murfreesboro, TN 37132, USA. [2]Department of Engineering Technology, Middle Tennessee State University, 1301 East Main Street, Murfreesboro, TN 37132, USA. ✉email: misa.faezipour@mtsu.edu

in brain tumor detection[14]. Visual Geometry Group 16-layered deep learning architecture (VGG16) is a CNN-based model known for its depth and great performance in image recognition, and is commonly used for brain tumor classification[15], leveraging transfer learning to improve accuracy. Despite VGG16's high classification performance, its vulnerability to adversarial attacks is a significant concern in image classification. These attacks involve subtle manipulations of input images that can lead to incorrect and potentially harmful predictions[16,17]. For example, a high-grade tumor could be misclassified as low-grade, resulting in inappropriate treatment. Defensive strategies such as adversarial training[16] and feature-squeezing[18] have been developed to mitigate this risk. Adversarial training involves retraining models using perturbed examples to improve robustness while feature-squeezing simplifies input data to reduce sensitivity to minor changes.

Adversarial defenses tailored for brain tumor classification remain sparse despite the crucial importance of robust and reliable models in medical imaging. Moreover, current defense strategies often focus on single-layer defenses, such as adversarial training or feature transformations, which may not provide sufficient robustness against sophisticated adversarial attacks like the Fast Gradient Sign Method (FGSM)[16] and Projected Gradient Descent (PGD)[17] attacks.

The novelty of this study lies in the development of a comprehensive multi-layered defense framework that integrates ensemble adversarial training (leveraging both FGSM and PGD attacks) with feature-squeezing techniques, specifically bit-depth reduction and Gaussian blurring within a unified pipeline. Unlike prior studies that typically adopt single-layer defenses, our approach systematically combines these strategies to enhance robustness without sacrificing computational efficiency. To the best of our knowledge, such an integrated, domain-specific defense mechanism tailored for MRI-based brain tumor classification has not been extensively explored, addressing a critical gap in the field of secure medical AI.

In this research, we employ a multi-layered defense strategy to enhance brain tumor classification models' robustness against adversarial attacks. Our research generates the following contributions:

1. *Systematic analysis of adversarial vulnerabilities in brain tumor classification:* We assess the susceptibility of the VGG16 model to adversarial attacks (FGSM and PGD), establishing a baseline for its performance without defenses. This highlights the need for robust defenses in medical AI.
2. *Implementation of a multi-layered defense mechanism:* We integrate adversarial training and feature-squeezing techniques (*bit-depth reduction and Gaussian blurring*) to create a stronger defense strategy, offering better protection compared to single-layer defenses.
3. *Evaluation through comprehensive simulations:* We employ two simulation strategies by varying FGSM and PGD attack parameters to evaluate defense performance, showing improved model robustness with post-defense accuracies of 54% (FGSM) and 47% (PGD).

Related work

Recent advancements in deep learning have significantly influenced medical image analysis, including brain tumor classification. However, these models remain vulnerable to adversarial attacks, which could undermine their reliability in clinical applications. This section examines two important areas of related work: (1) the vulnerability of medical imaging models to adversarial attacks and (2) defense mechanisms developed to mitigate these vulnerabilities. Finally, we outline how this study addresses existing gaps in these areas.

Vulnerability to adversarial attacks in medical imaging

Adversarial attacks expose deep learning models' susceptibility to minor input perturbations, raising concerns about their reliability in essential applications. This vulnerability was initially highlighted by Goodfellow et al.[16] through FGSM and later expanded by Madry et al.[17] with PGD method, which exposed how slight perturbations can lead to significant misclassifications. Recent studies[19–29] confirmed that this vulnerability raises significant concerns in medical applications, where misclassifications could have severe clinical consequences. In brain tumor classification, specific vulnerabilities to adversarial attacks have been identified[30]. Similarly, Joel et al.[31] systematically reviewed vulnerabilities in medical imaging models, showing that minor perturbations in MRI images can significantly impact diagnostic accuracy. This growing body of work highlights the urgent need for robust defense mechanisms in medical imaging to mitigate the risk posed by adversarial attacks[32].

Defense mechanisms against adversarial attacks

To counter adversarial vulnerabilities, researchers have proposed several defense strategies, including adversarial training[33] and feature-squeezing[18]. Adversarial training[16] involves training models with adversarial examples to improve robustness against adversarial attacks such as FGSM and PGD attacks. Madry et al.[17] developed this approach with PGD-based adversarial training, providing resilience against stronger adversarial perturbations. Moreover, Pang et al.[34] explored hybrid defense strategies that integrate adversarial training with other techniques to enhance model robustness. Feature-squeezing[18] reduces the model's sensitivity to noise by decreasing image bit-depth and applying Gaussian blurring, thereby limiting the degree of freedom for adversarial noise.

Limitations in existing research

Although previous studies[35–39] have achieved high classification accuracy and explored isolated defense mechanisms, most do not provide comprehensive protection against multiple adversarial attacks. Isolated defense strategies may be effective against specific attacks but often fall short in generalizability, particularly in high-risk, real-world applications such as medical imaging. Additionally, as Pang et al.[34] explored, ensemble defense strategies often require significant computational resources, posing challenges for practical application in time-sensitive medical scenarios. Hence, our approach addresses the limitations of single-layer defenses by integrating techniques such as adversarial training and feature-squeezing, ensuring both effectiveness and computational efficiency.

Methods

In this research, we employed a deep learning model using VGG16 (CCN-based architecture) to classify brain MRI images into four categories: *glioma, meningioma, pituitary tumor*, and *no tumor*. Given the unique challenges of medical imaging, particularly the susceptibility of models to adversarial attacks, we incorporated various defense mechanisms to improve model robustness. This section explains the model architecture and the adversarial attack techniques employed in the research. Figure 1 depicts the research pipeline for the study.

Model architecture

The VGG16 is pre-trained on the ImageNet dataset[40] and adapted for the brain MRI classification through transfer learning in this research. Serving as the feature extraction backbone, its fully connected layers are replaced with a custom classifier for image classification. The modified architecture includes Global Average Pooling (GAP) to reduce feature dimensions, a dense layer with 128 neurons activated by ReLU, and a dropout layer (*rate 0.5*) to prevent overfitting. The final output layer, with 4 neurons and softmax activation, generates probability scores for each class. We compile the model with the Adam optimizer (*learning rate* of 10^{-3}), using sparse categorical cross-entropy as the loss function and accuracy as the primary evaluation metric. Moreover, we chose VGG16 in this research because of its simplicity, proven effectiveness in medical imaging tasks, and compatibility with transfer learning. VGG16's sequential, uniform architecture allows for easier interpretability and modification, particularly advantageous when integrating defense mechanisms like adversarial training and feature-squeezing. Its depth and well-structured layers enable effective feature extraction from MRI images, which is crucial for capturing the nuanced differences in brain tumor types. Additionally, VGG16 achieves strong classification performance while maintaining computational efficiency, making it an ideal foundation for developing robust, explainable, and secure medical imaging models.

Adversarial attack techniques

We employ two adversarial attack methods, the Fast Gradient Sign Method and Projected Gradient Descent, to test our model's resilience against adversarial perturbations. FGSM is a quick, *single-step* attack that perturbs the input image in the gradient direction, increasing the model's prediction error with minimal modifications. PGD, a more robust *iterative* attack, refines this process by applying *multiple small steps* within a bounded region, creating stronger adversarial examples. Additionally, we apply feature-squeezing techniques (*bit-depth reduction* and *Gaussian blurring*) to mitigate adversarial effects by reducing image sensitivity to minor perturbations. Together, these techniques allow us to evaluate and improve model robustness rigorously. Figure 2 shows the adversarial attack pipeline before adversarial training.

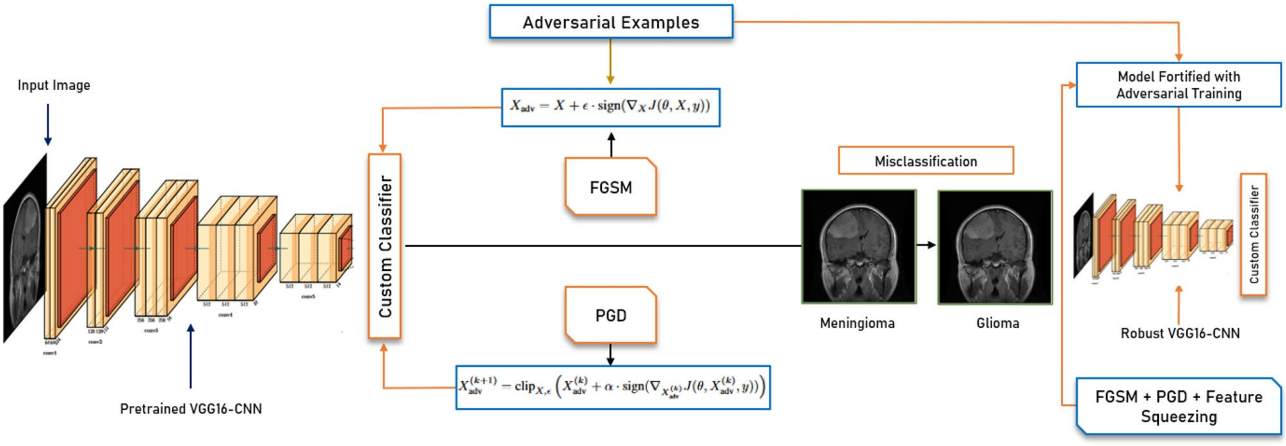
Adversarial Examples
Input Image
$X_{\text{adv}} = X + \epsilon \cdot \text{sign}(\nabla_X J(\theta, X, y))$
Model Fortified with Adversarial Training
FGSM
Misclassification
Custom Classifier
Custom Classifier
Meningioma
Glioma
Robust VGG16-CNN
PGD
Pretrained VGG16-CNN
$X_{\text{adv}}^{(k+1)} = \text{clip}_{X,\epsilon}\left(X_{\text{adv}}^{(k)} + \alpha \cdot \text{sign}(\nabla_{X_{\text{adv}}^{(k)}} J(\theta, X_{\text{adv}}^{(k)}, y))\right)$
FGSM + PGD + Feature Squeezing

Fig. 1. Research pipeline for adversarial training and defense mechanism in brain tumor classification. The pipeline begins with processing an input MRI image by a pre-trained VGG16-CNN for feature extraction and adding a custom classifier for the brain tumor classification. Afterward, adversarial examples are generated using **FGSM** and **PGD** attacks, introducing perturbations leading to potential misclassifications (e.g., a *meningioma* image being classified as *glioma*). The adversarial examples are integrated into the training phase alongside feature-squeezing techniques, resulting in a robust VGG16-CNN model fortified against adversarial attacks. This multi-layered approach ensures improved resilience for accurate brain tumor classification in crucial medical imaging.

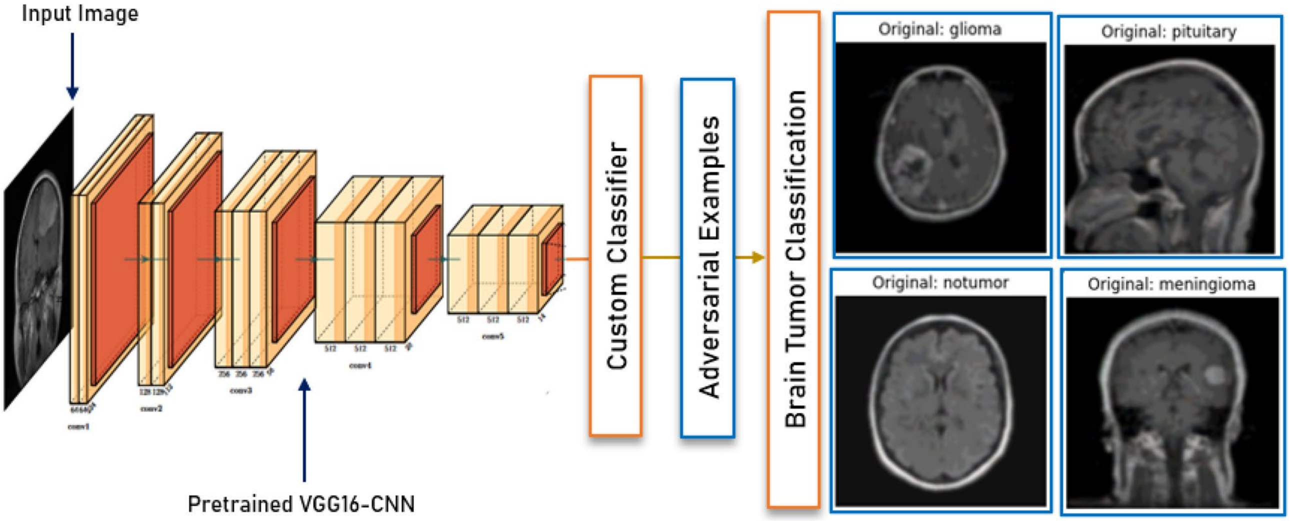

Fig. 2. Brain tumor classification pipeline with adversarial examples. The model employs a *pre-trained VGG16-CNN* as a **feature extractor**, followed by a **custom classifier** tailored for brain tumor detection. *Adversarial examples*, generated to assess model robustness, are incorporated to challenge the model's accuracy across four categories: **glioma**, **pituitary**, **notumor**, and **meningioma**. This setup evaluates the model's resilience and effectiveness in *medical imaging*.

```
Require: Model $f$, image $X$, true label $y$, FGSM perturbation $\varepsilon_{\text{FGSM}}$, PGD parameters $\varepsilon_{\text{PGD}}$, $\alpha$, iterations $T$, bit-depth $b$, Gaussian kernel size $k$
Ensure: Adversarially trained image $X_{\text{adv\_train}}$
Initialize $X_{\text{adv\_train}} = X$
feature-squeezing for Defense:
Reduce bit-depth of $X$: $X_{\text{reduced}} = \text{round}(X \times (2^b - 1))/(2^b - 1)$
Apply Gaussian blur: $X_{\text{squeezed}} = X_{\text{reduced}} * G(k)$
FGSM Adversarial Image Generation:
Compute gradient of the loss w.r.t input $X_{\text{squeezed}}$: $g_{\text{FGSM}} = \nabla_{X_{\text{squeezed}}} J(\theta, X_{\text{squeezed}}, y)$
Calculate FGSM perturbation: $\delta_{\text{FGSM}} = \varepsilon_{\text{FGSM}} \cdot \text{sign}(g_{\text{FGSM}})$
Generate FGSM adversarial image: $X_{\text{adv\_FGSM}} = X_{\text{squeezed}} + \delta_{\text{FGSM}}$
Clip $X_{\text{adv\_FGSM}}$ to ensure valid pixel values (e.g., [0, 1] range)
PGD Adversarial Image Generation:
Initialize $X_{\text{adv\_PGD}} = X_{\text{squeezed}}$
for $k = 1$ to $T$ do
    Compute gradient of the loss w.r.t $X_{\text{adv\_PGD}}$: $g_{\text{PGD}} = \nabla_{X_{\text{adv\_PGD}}} J(\theta, X_{\text{adv\_PGD}}, y)$
    Update PGD adversarial image: $X_{\text{adv\_PGD}} = X_{\text{adv\_PGD}} + \alpha \cdot \text{sign}(g_{\text{PGD}})$
    Project $X_{\text{adv\_PGD}}$ back into $\varepsilon_{\text{PGD}}$-ball around $X_{\text{squeezed}}$
    Clip $X_{\text{adv\_PGD}}$ to maintain valid pixel values
end for
Combine Adversarial Examples for Training:
Set $X_{\text{adv\_train}} = X_{\text{adv\_FGSM}}$ or $X_{\text{adv\_PGD}}$ based on the chosen adversarial strategy
return $X_{\text{adv\_train}}$
```

Algorithm 1. Adversarial Training with feature-squeezing, FGSM, and PGD

Algorithm 1 explains the process of the multi-layered strategy (*integrated defense framework*) in this paper using the feature-squeezing techniques with the FGSM and PGD attacks in the adversarial training procedure to improve the model robustness. Initially, feature-squeezing reduces bit-depth (b) and applies Gaussian blurring with kernel size k, which minimizes sensitivity to minor input perturbations. FGSM then generates adversarial examples by calculating perturbations using ϵ_{FGSM}. At the same time, PGD further refines these adversarial images through iterative updates based on ϵ_{PGD}, step size α, and iterations T. The model is subsequently trained on these adversarial examples, enhancing resilience against adversarial attacks.

Component	Specification
Platform	Google Colab Pro+
Processor (CPU)	Intel Xeon CPU @ 2.20GHz (Colab environment)
Graphics processing unit (GPU)	NVIDIA A100-SXM4-40GB (Google Colab)
RAM (Colab)	80 GB
Local system	Windows 10, Intel Core i5 ThinkPad, 8GB RAM
Programming language	Python 3.9
Deep learning framework	TensorFlow 2.11.0, Keras API
CUDA version	CUDA Toolkit 11.2 (Colab Environment)
Dataset	Composite MRI dataset (7,023 images), combining Figshare, SARTAJ, and Br35H datasets
Image preprocessing	Image resizing to 128x128 pixels, normalization (0-1 range)
Model architecture	Pre-trained VGG16 with custom classifier layers
Optimizer	Adam Optimizer (learning rate: 10^{-3})
Loss function	Sparse categorical cross-entropy
Evaluation metrics	Accuracy, Precision, Recall, F1-Score
Defense techniques	Adversarial Training (FGSM, PGD), Feature Squeezing (bit-depth reduction, Gaussian blurring)

Table 1. Experimental Setup: Hardware and Software Configurations.

Class	Precision	Recall	F1-Score	Support
Pituitary	1.00	0.95	0.97	292
Glioma	0.95	0.95	0.95	266
Meningioma	0.91	0.95	0.93	273
No tumor	0.97	0.99	0.98	312
Accuracy	0.96			
Macro Avg	0.96	0.96	0.96	1143
Weighted Avg	0.96	0.96	0.96	1143

Table 2. Initial classification report.

Experimental setup

Dataset preparation

We employed a composite dataset[41] sourced from Kaggle, combining three datasets: Figshare[42], SARTAJ[43], and Br35H[44]. This collection includes 7,023 MRI brain images, categorized into four classes: *glioma, meningioma, pituitary tumor*, and *no tumor*. Gliomas originate from glial cells, meningiomas develop from the meninges, and pituitary tumors arise on the pituitary gland at the brain's base[45]. The "*no tumor*" category consists of MRI images without visible tumors, sourced exclusively from the Br35H dataset, while the remaining tumor images are drawn from the SARTAJ and Figshare datasets. This diverse dataset provides a comprehensive foundation for evaluating brain tumor classification models across multiple tumor types.

Hardware and software configuration

The experiments were conducted using the following hardware and software configurations in Table 1:

Implementation details

In this section, we focused on resizing the MRI images to 128 x 128 pixels with normalized pixel values and then split the dataset with an 80:20 ratio for training and validation. Initially, we trained the model for 5 epochs with VGG16's layers frozen to establish a baseline. For fine-tuning, we unfreezed the last 10 layers, using a lower learning rate of 10^{-3} for an additional 10 epochs, achieving strong validation performance for further testing with adversarial defenses.

Experiments and results

Baseline performance

We trained the model on the clean brain tumor MRI dataset before introducing adversarial examples to test its robustness. The initial model achieved a high classification accuracy, validating its effectiveness in distinguishing between *glioma, meningioma, pituitary tumor*, and *no tumor classes*.

Results

Table 2 depicts the initial classification report of the model showing a higher performance with an overall accuracy of **96%** across four brain tumor classes. Precision, recall, and F1 scores are consistently high, particularly for the

Attack type	Class	Pre-adversarial training			Post-adversarial training		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
FGSM	Pituitary	0.19	0.06	0.09	**0.58**	**0.73**	**0.65**
	Glioma	0.30	0.23	0.26	**0.57**	**0.53**	**0.55**
	Meningioma	0.14	0.29	0.19	0.24	0.16	0.19
	No tumor	0.71	0.63	0.67	**0.75**	**0.82**	**0.78**
PGD	Pituitary	0.01	0.00	0.01	**0.47**	**0.54**	**0.50**
	Glioma	0.04	0.03	0.03	0.30	0.20	0.24
	Meningioma	0.04	0.09	0.06	0.24	0.17	0.20
	No Tumor	0.50	0.35	0.41	**0.64**	**0.89**	**0.75**

Table 3. Classification Report on FGSM and PGD Adversarial Data Before and After Adversarial Training. The **bolded** values in the *Post-Adversarial Training* columns emphasize significant improvements in **Precision**, **Recall**, and **F1-Score** for specific classes, showcasing enhanced model resilience against both *FGSM* and *PGD* attacks after adversarial defenses were applied. These improvements are most notable in the *Pituitary* and *No Tumor* classes, indicating that the adversarial training effectively reinforced the model's ability to withstand perturbations.

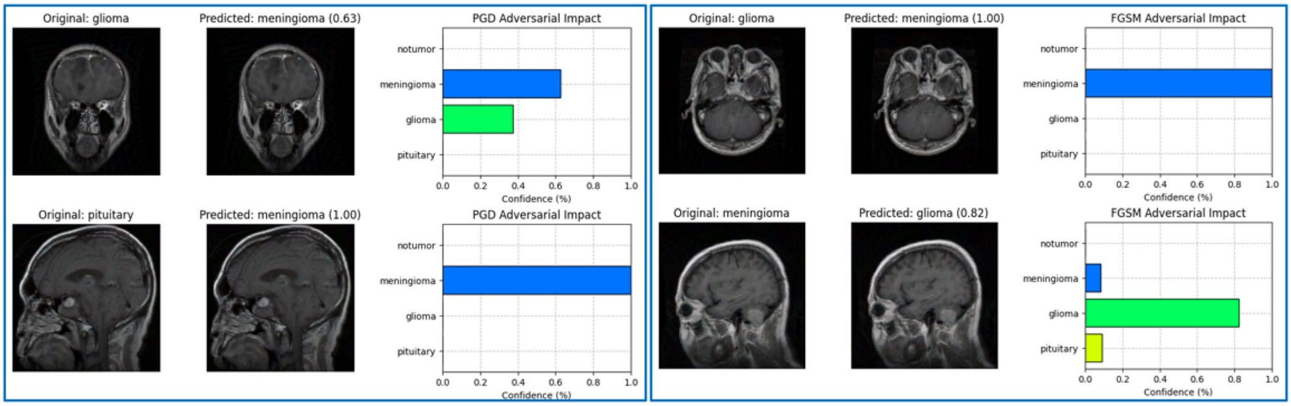

Fig. 3. **Confidence bar plots** showing the impact of **FGSM** and **PGD** adversarial attacks on the model's classification accuracy. In the ***PGD attack*** (left), the model misclassifies a *glioma* image as *meningioma* with a moderate confidence of 0.63. In contrast, a *pituitary* image is classified as *meningioma* with high certainty. In the ***FGSM attack*** (right), the model shows high confidence (1.00) in misclassifying a *glioma* image as *meningioma* and a confidence of 0.82 in classifying a *meningioma* image as *glioma*. These results emphasize the model's vulnerability to adversarial perturbations, leading to incorrect yet high-confidence predictions.

Pituitary and No Tumor classes. This robust baseline establishes a solid foundation for assessing the impact of adversarial defenses.

Adversarial attack evaluation (pre-defense)

1. **FGSM Attack** (*Pre-Adversarial Training*): We applied the FGSM attack to the model before adversarial training, using a parameter value of $\epsilon = 0.01$. This attack reduced the model's accuracy on FGSM-generated adversarial data to 32%, highlighting its initial vulnerability to adversarial perturbations before implementing defense mechanisms.
2. **PGD Attack** (*Pre-Adversarial Training*): We conducted the PGD attack on the model before implementing adversarial training. For this attack, we set the parameters to $\epsilon = 0.01$, $\alpha = 0.002$, and utilized 10 iterations. This pre-adversarial training attack significantly reduced the model's accuracy on PGD adversarial data to 13%, further showing the model's susceptibility to adversarial examples.

Table 3 depicts the classification report that shows the model's vulnerability to FGSM and PGD attacks on the **Pre-Adversarial Training** column on the table. While No Tumor images retain moderate precision and recall under both attacks, tumor classes like Pituitary, Glioma, and Meningioma exhibit significant performance degradation, especially under PGD. This highlights the need for enhanced adversarial defenses.

Moreover, Fig. 3 also shows the combined confidence bar plots displaying the model's responses to both FGSM and PGD adversarial examples

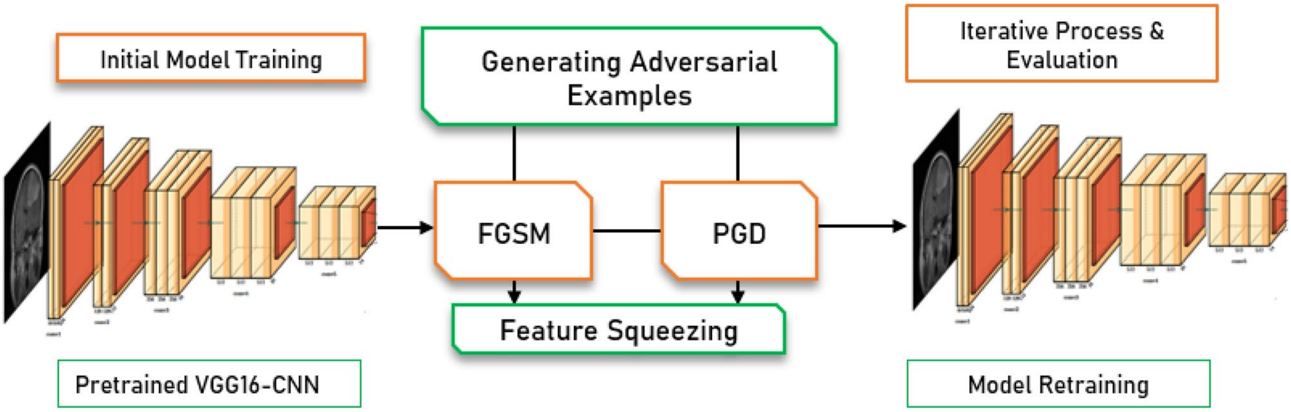

Fig. 4. Pipeline for adversarial training and model robustness enhancement using a pre-trained VGG16 for brain tumor classification. The process starts with *Initial Model Training* on standard MRI images, followed by the generation of adversarial examples using **FGSM** and **PGD** methods. **feature-squeezing** is applied to mitigate adversarial impacts, and an *Iterative Process* is employed for model retraining and evaluation. This iterative approach aims to improve the model's resilience against adversarial attacks.

Simulation	**FGSMϵ**	**PGD Iter.**	**PGDϵ**	**PGDα**
First Simulation	0.05	10	0.05	0.04
Second Simulation	0.01	10	0.01	0.002
feature-squeezing Parameters				
Bit-depth: 4 bits, Gaussian Blurring: Kernel size 3 × 3, adjusted std.				

Table 4. Simulation parameters for adversarial training and feature-squeezing.

Adversarial training and defense mechanisms

In this study, Fig. 4 shows the adversarial training pipeline for the defense mechanism for the model. We implemented the adversarial training procedure by augmenting and incorporating FGSM and PGD adversarial examples alongside feature-squeezing techniques to enhance the model's robustness and resilience against attacks by counteracting subtle input perturbations, including 4-bit depth reduction and Gaussian blurring (**kernel size** 3×3). This approach involved two simulation strategies, where only specific parameters within the FGSM and PGD functions in Algorithm 1, such as the ϵ values for FGSM and the ϵ and α values for PGD, were varied. All other values, including those used in feature-squeezing and the process for generating adversarial examples, remained constant throughout both simulations. This controlled adjustment of parameters was intended to effectively assess and optimize the model's resilience to adversarial perturbations.

Simulation 1: baseline defense strategy

In the first simulation, we set FGSM $\epsilon = 0.05$ and configured PGD with parameters $\epsilon = 0.05$, $\alpha = 0.04$, and 10 iterations as shown in Table 4 . These moderate perturbation levels were selected to evaluate the model's initial resilience and defensive capacity. The adversarial training was conducted over 10 epochs on a combined dataset of clean and adversarially perturbed images, with consistent feature-squeezing applied.

Simulation 2: enhanced defense strategy

In the second simulation, we adjusted the FGSM $\epsilon = 0.01$ and PGD parameters to $\epsilon = 0.01$, $\alpha = 0.002$, with 10 iteration as shown in Table 4. This lower perturbation threshold was employed to simulate less aggressive attacks and test the model's robustness under moderate adversarial conditions. Feature-squeezing settings remained unchanged to provide a controlled comparison across simulations.

Ablation studies

To evaluate the specific impact of parameter changes on model robustness, we conducted an ablation study focusing on the FGSM and PGD parameters alone. By varying these parameters while keeping feature-squeezing and other configurations constant, we aimed to isolate the effect of each parameter set on adversarial resilience. This analysis allowed us to determine the optimal configuration of adversarial training parameters for maximum robustness. The main findings include:

- **FGSM Parameter Impact:** Higher ϵ values led to more aggressive perturbations, causing a noticeable drop in model accuracy before defense mechanisms were applied. However, adversarial training with a moderate ϵ(0.05 in *Simulation* 1) provided a balanced trade-off between attack strength and model resilience.

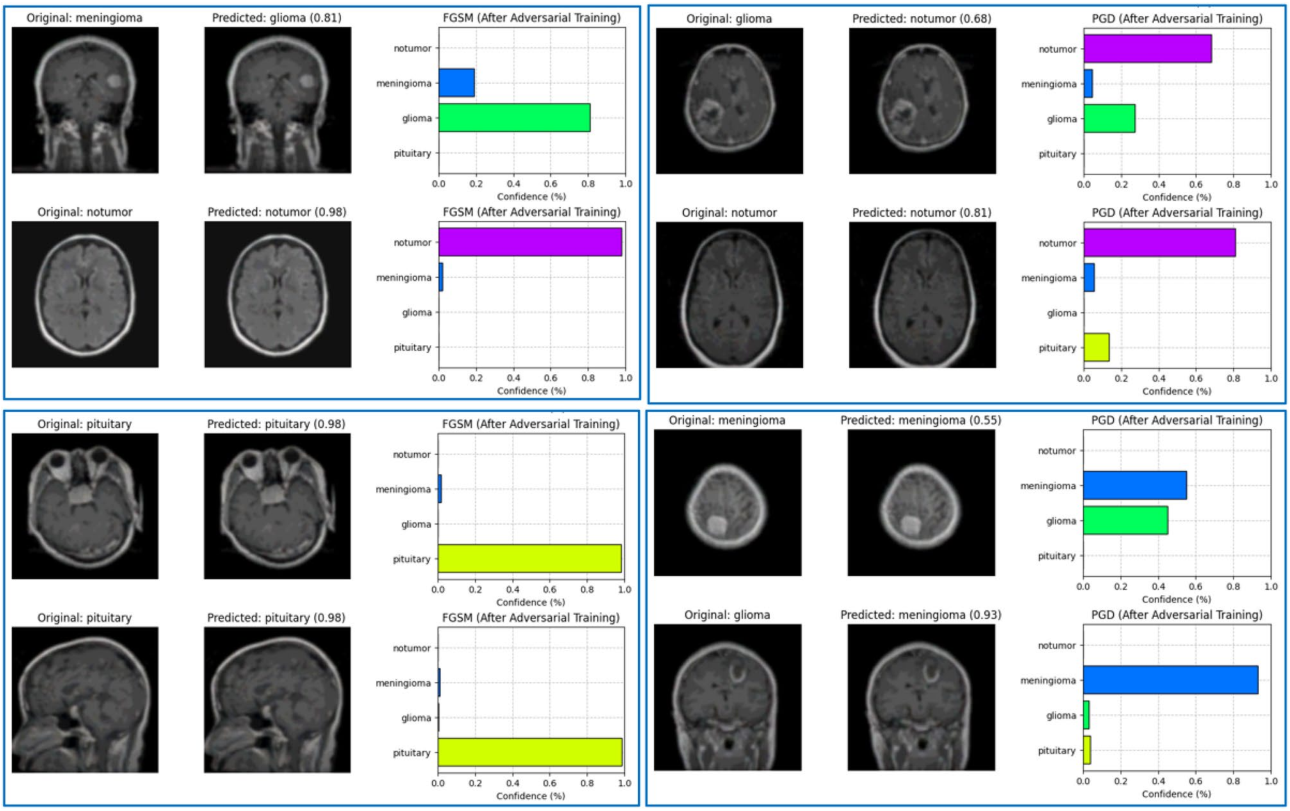

Fig. 5. Post-adversarial training evaluation on FGSM and PGD attacks across two simulations. The figure shows confidence bar plots for model predictions following adversarial training with **FGSM** and **PGD** adversarial examples. **Left**: FGSM impact in both simulations, where the model shows high confidence in correct classifications like “no tumor” and “pituitary” post-defense. **Right**: PGD impact, with improved resilience as the model maintains correct classifications with balanced confidence levels. This visualization highlights the efficacy of combined defense mechanisms in enhancing model robustness against adversarial attacks.

- **PGD Parameter Impact:** The step size α and the number of iterations significantly influenced the model’s performance under adversarial conditions. A lower α and ϵ in PGD (as in *Simulation* 2) introduced a more gradual perturbation process, making it easier for the model to classify images correctly after adversarial training. This setup demonstrated a controlled increase in robustness without sacrificing performance on clean images.

Results: post-adversarial training evaluation on adversarial data

After the adversarial training and feature-squeezing, we comprehensively evaluated FGSM and PGD adversarial examples to measure the model’s enhanced resilience. To provide a clear summary of the effectiveness of our defense mechanisms, we report the model’s prediction robustness employing confidence bar plots in Fig. 5 for the two simulations, focusing on the overall improvement rather than individual simulation results.

Discussion

In this research, we examined the effectiveness of a multi-layered defense strategy to improve the robustness of a brain tumor classification model against adversarial attacks. We aimed to reduce the model’s vulnerability to subtle but impactful adversarial perturbations by employing an adversarial training defensive mechanism, which integrates FGSM and PGD adversarial examples alongside feature-squeezing techniques such as bit depth reduction and Gaussian blurring in the defensive training process. We performed experiments using two simulation strategies with different FGSM and PGD parameters to systematically assess and optimize the model’s resilience. Our results demonstrate that carefully calibrated adversarial training parameters can achieve significant performance on adversarially perturbed data without compromising accuracy on clean data.

Moreover, the confidence bar plots illustrate how our defense strategy influenced the model’s confidence distributions post-adversarial training, displaying balanced and reduced (*optimized*) confidence in misclassified adversarial examples. This change shows that the model was now more accurate and cautious in the presence of subtle and challenging inputs. We effectively reduced the effects of minor perturbations that adversarial attacks exploit by employing feature-squeezing. The adversarial training and the feature-squeezing’s simplified input representation made it difficult for adversarial perturbations to deceive the model.

It is also important to note that adversarial perturbations frequently result in high-confidence misclassifications, as illustrated in Fig. 3. These perturbations push inputs across decision boundaries while maintaining strong softmax outputs, leading to overconfident incorrect predictions. Post-processing techniques such as temperature scaling or Platt scaling can be applied to calibrate the model's output probabilities and reduce overconfidence to mitigate this issue. Integrating such confidence threshold adjustments may further enhance model reliability, especially in critical clinical decision-making scenarios.

Furthermore, scaling adversarial training for larger datasets or real-time clinical applications presents additional challenges due to increased computational demands. To address this, several optimization strategies can be considered in future implementations. Techniques such as model quantization, pruning, and knowledge distillation can significantly reduce model complexity while preserving robustness. Leveraging lightweight architectures and limiting the number of adversarial examples during training can also enhance efficiency without sacrificing defense performance. Incorporating these approaches would make the proposed defense framework more adaptable for real-time, resource-constrained clinical environments.

Clinical applicability

The proposed multi-layered defense strategy holds significant potential for real-world clinical integration. Specifically, the adversarially trained model can be embedded within AI-assisted diagnostic systems used in hospitals and healthcare facilities. Integration into Picture Archiving and Communication Systems (PACS) or cloud-based medical AI platforms can provide clinicians with enhanced diagnostic reliability by ensuring the model remains robust against potential adversarial manipulations. Furthermore, this defense framework can serve as an additional verification layer, safeguarding against incorrect classifications and supporting safer clinical decision-making processes.

Limitations and future work

Despite the encouraging results, several limitations should be acknowledged. The integration of multiple defense techniques, specifically adversarial training combined with feature-squeezing, introduces additional computational overhead, which could impact scalability when applied to larger datasets or more complex architectures. The increased training time and resource requirements may pose challenges in real-time clinical implementations. Additionally, this study primarily evaluates the defense framework on the VGG16 architecture and focuses on two commonly used adversarial attack methods (FGSM and PGD). Expanding this investigation to include other neural network architectures and a wider variety of attack techniques, such as Carlini-Wagner or DeepFool attacks, could further validate the robustness and generalizability of the proposed defense mechanism. Future work will also explore optimizing the computational efficiency of the defense strategy and extending the framework to other medical imaging modalities beyond MRI-based brain tumor classification. Investigating adaptive defense mechanisms tailored for real-time clinical environments represents another promising direction to ensure reliable AI-assisted diagnostics under adversarial conditions. Furthermore, exploring ensemble-based defense strategies, such as combining multiple models or classifiers, may offer additional robustness against adversarial attacks. Incorporating such approaches will be considered in future work to enhance the overall defense capability while balancing computational efficiency. In addition, integrating complementary defense strategies such as adversarial detection mechanisms and alternative input preprocessing techniques, including wavelet filtering or noise reduction methods, may further strengthen model resilience. These avenues will be explored in subsequent research to improve the robustness and reliability of AI-driven medical imaging systems.

Conclusion

This study presents a multi-layered defense strategy to enhance the robustness of a CNN-based brain tumor classification model against adversarial attacks. We addressed crucial challenges in the medical image (*brain tumor*) analysis and the model's vulnerability to adversarial perturbations by employing adversarial training techniques with FGSM and PGD attacks and utilizing feature-squeezing techniques to fortify this model. We fine-tuned the model using adversarially perturbed examples using the pre-trained VGG16 architecture and feature-squeezing to reduce the impact of subtle perturbations. These processes (*modifications*) enabled us to significantly improve accuracy on adversarial data while maintaining robust performance on clean data. Our experiments across two simulation strategies showed that the model adapts to different adversarial attacks and intensities by varying the FGSM and PGD parameter perturbation magnitudes. Furthermore, our results highlight the effectiveness of our approach, which is enhanced by including the feature-squeezing techniques with the adversarial training procedure. The model's post-adversarial training evaluation demonstrated marked improvement in accuracy and confidence distributions, suggesting that it became more accurate and highly reliable when dealing with adversarial inputs.

For medical applications, where adversarial attacks on AI medical models resulting in erroneous diagnostic evaluation are consequentially severe, more robust models can help towards safer and more reliable AI-assisted diagnosis. Our research provides a scalable and adaptable framework for adversary robustness in healthcare and other high-risk domains where reliability and dependability on AI-powered systems are imperative.

Data availibility

The datasets generated and/or analyzed during the current study are available in the Kaggle repository (Brain Tumor MRI Dataset).

Received: 31 January 2025; Accepted: 2 May 2025
Published online: 14 May 2025

References

1. Villanueva-Meyer, J. E., Mabray, M. C. & Cha, S. Current clinical brain tumor imaging. *Neurosurgery* **81**(3), 397–415. https://doi.org/10.1093/neuros/nyx103 (2017).
2. Ostrom, Q. T. et al. CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2015–2019. *Neuro-Oncology* **24**, 1. https://doi.org/10.1093/neuonc/noac202 (2022).
3. Abdusalomov, A. B., Mukhiddinov, M. & Whangbo, T. K. Brain tumor detection based on deep learning approaches and magnetic resonance imaging. *Cancers* **15**(16), 4172–4172. https://doi.org/10.3390/cancers15164172 (2023).
4. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Ann. Rev. Biomed. Eng.* **19**(1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442 (2017).
5. Naqvi, M., Gilani, S. Q., Marques, O. & Kim, H.-C. Skin cancer detection using deep learning—A review. *Diagnostics* **13**(11), 1911–1911. https://doi.org/10.3390/diagnostics13111911 (2023).
6. Bhimavarapu, U. & Battineni, G. Deep learning for the detection and classification of diabetic retinopathy with an improved activation function. *Healthcare* **11**(1), 97. https://doi.org/10.3390/healthcare11010097 (2022).
7. Mahmood, T., Rehman, A., Saba, T., Wang, Y. & Alamri, F. S. Alzheimer's disease unveiled: Cutting-edge multi-modal neuroimaging and computational methods for enhanced diagnosis. *Biomed. Signal Process. Control* **97**, 106721. https://doi.org/10.1016/j.bspc.2024.106721 (2024).
8. Mahmood, T., Saba, T., Rehman, A. & Alamri, F. S. Harnessing the power of radiomics and deep learning for improved breast cancer diagnosis with multiparametric breast mammography. *Expert Syst. Appl.* **249**, 123747. https://doi.org/10.1016/j.eswa.2024.123747 (2024).
9. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**(7639), 115–118. https://doi.org/10.1038/nature21056 (2017).
10. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**(1), 60–88. https://doi.org/10.1016/j.media.2017.07.005 (2017).
11. Sarvamangala, D. R. & Kulkarni, R. V. Convolutional neural networks in medical image understanding: A survey. *Evolut. Intell.* https://doi.org/10.1007/s12065-020-00540-3 (2021).
12. Ullah, N. et al. An effective approach to detect and identify brain tumors using transfer learning. *Appl. Sci.* **12**(11), 5645. https://doi.org/10.3390/app12115645 (2022).
13. Ullah, N., Khan, M. S., Khan, J. A., Choi, A. & Anwar, M. S. A robust end-to-end deep learning-based approach for effective and reliable BTD using MR images. *Sensors* **22**(19), 7575. https://doi.org/10.3390/s22197575 (2022).
14. Ullah, N., Hassan, M., Khan, J. A., Anwar, M. S. & Aurangzeb, Khursheed. Enhancing explainability in brain tumor detection: A novel DeepEBTDNet model with LIME on MRI images. *Int. J. Imaging Syst. Technol.* https://doi.org/10.1002/ima.23012 (2023).
15. Simonyan, K., & Zisserman, A. Very deep convolutional networks for large-scale image recognition (2015). arXiv:abs/1409.1556
16. Goodfellow, I.J., Shlens, J., & Szegedy, C. Explaining and harnessing adversarial examples (2014). arXiv:abs/1412.6572
17. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. Towards deep learning models resistant to adversarial attacks (2019). arXiv:abs/1706.06083
18. Xu, W., Evans, D., & Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings 2018 Network and Distributed System Security Symposium* (2018). https://doi.org/10.14722/ndss.2018.23198
19. Finlayson, S. G. et al. Adversarial attacks on medical machine learning. *Science* **363**(6433), 1287–1289. https://doi.org/10.1126/science.aaw4399 (2019).
20. Kotia, J., Kotwal, A., & Bharti, R. Risk susceptibility of brain tumor classification to adversarial attacks. In *Proceedings of International Conference on Advances in Computing and Data Sciences*, 193–201 (2019). https://doi.org/10.1007/978-3-030-31964-9_17
21. Ma, X. et al. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognit.* https://doi.org/10.1016/j.patcog.2020.107332 (2020).
22. Paul, R. *et al.*, Mitigating adversarial attacks on medical image understanding systems. In *ISBI 2020* (2020). https://doi.org/10.1109/isbi45749.2020.9098740
23. Paschali, M., Bian, W., Shi, W., Wang, C. & Rueckert, D. Generalizability versus robustness: Adversarial examples for medical imaging. *Med. Image Anal.* **72**, 102178. https://doi.org/10.1016/j.media.2021.102178 (2021).
24. Apostolidis, K. D. & Papakostas, G. A. A survey on adversarial deep learning robustness in medical image analysis. *Electronics* **10**(17), 2132. https://doi.org/10.3390/electronics10172132 (2021).
25. Kaviani, S., Han, K. J. & Sohn, I. Adversarial attacks and defenses on AI in medical imaging informatics: A survey. *Expert Syst. Appl.* **198**, 116815. https://doi.org/10.1016/j.eswa.2022.116815 (2022).
26. Rodriguez, D. et al. On the role of deep learning model complexity in adversarial robustness for medical images. *BMC Med. Inform. Decis. Mak.* **22**, 2. https://doi.org/10.1186/s12911-022-01891-w (2022).
27. Shi, X. et al. Robust convolutional neural networks against adversarial attacks on medical images. *Pattern Recognit.* **132**, 108923. https://doi.org/10.1016/j.patcog.2022.108923 (2022).
28. Tsai, M.-J., Lin, P.-Y. & Lee, M.-E. Adversarial attacks on medical image classification. *Cancers* **15**(17), 4228. https://doi.org/10.3390/cancers15174228 (2023).
29. Muoka, G. W. et al. A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense. *Mathematics* **11**(20), 4272. https://doi.org/10.3390/math11204272 (2023).
30. Ma, S., Mathur, P., Ju, Z., Lawlor, A. & Dong, R. Model-data-driven adversarial active learning for brain tumor segmentation. *Comput. Biol. Med.* **176**, 108585. https://doi.org/10.1016/j.compbiomed.2024.108585 (2024).
31. Joel, M. Z. et al. Using adversarial images to assess the robustness of deep learning models trained on diagnostic images in oncology. *JCO Clin. Cancer Inform.* **6**, e2100170. https://doi.org/10.1200/CCI.21.00170 (2022).
32. Sheikh, Z. & Zafar, A. Robust medical diagnosis: A novel two-phase deep learning framework for adversarial proof disease detection in radiology images. *J. Imaging Inform. Med.* **37**(1), 308–338. https://doi.org/10.1007/s10278-023-00916-8 (2024).
33. Wu, B. *et al.*, Defenses in adversarial machine learning: A survey (2023). arXiv:pdf/2312.08890
34. Pang, T., Xu, K., Du, C., Chen, N., & Zhu, J. Improving adversarial robustness via promoting ensemble diversity (2019). arXiv:abs/1901.08846
35. Macas, M., Wu, C. & Fuertes, W. Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems. *Expert Syst. Appl.* **238**, 122223. https://doi.org/10.1016/j.eswa.2023.122223 (2024).
36. Sen, J., & Dasgupta, S. Adversarial attacks on image classification models: FGSM and patch attacks and their impact (2023). arXiv:abs/2307.02055
37. Cai, Z., Tan, Y., & Salman, A.M. Ensemble-based blackbox attacks on dense prediction (2023). arXiv:abs/2303.14304
38. Huang, S., Lu, Z., Deb, K., & Boddeti, V. N. Revisiting residual networks for adversarial robustness: An architectural perspective (2022). arXiv:abs/2212.11005
39. Huang, B. *et al.* Boosting accuracy and robustness of student models via adaptive adversarial distillation. In *CVPR 2023*, (2023). https://openaccess.thecvf.com/content/CVPR2023/papers/Huang_Boosting_Accuracy_and_Robustness_of_Student_Models_via_Adaptive_Adversarial_CVPR_2023_paper.pdf
40. *ImageNet*. https://www.image-net.org/
41. Nickparvar, M. "Brain tumor MRI dataset," (2021). https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset
42. *Brain tumor dataset*, (2021). https://doi.org/10.6084/m9.figshare.1512427.v5

43. *Brain Tumor Classification (MRI)*, (2021). https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri
44. *Br35H: : Brain Tumor Detection 2020*, (2020). https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection?select=no
45. *Brain Tumor Types*. https://www.hopkinsmedicine.org/health/conditions-and-diseases/brain-tumor/brain-tumor-types

Author contributions

Conceptualization, A.Y. and M.F.; Methodology, A.Y. and M.F.; Data curation, Visualization, A.Y. and M.F.; Formal analysis, A.Y. and M.F.; Software, A.Y. and M.F.; Validation, A.Y. and M.F.; Writing - original draft, A.Y. and M.F.; Writing - Review and Editing, A.Y. and M.F; Supervision, M.F.; Project administration, M.F.

Funding

No external funding was received for this study.

Declarations

Competing interests

The authors declare no competing interests.

Ethical considerations

The datasets used in this study were obtained from publicly available sources, specifically the Figshare[42], SARTAJ[43], Br35H[44], and Nickparvar[41] datasets. All these datasets are hosted on Kaggle or open-access repositories and are openly accessible. These datasets are fully anonymized and contain no personally identifiable patient information. As the data were de-identified prior to public release, there was no requirement for additional ethical approval or patient consent for their use in this research. The study adheres to ethical standards for data privacy and complies with relevant guidelines regarding the use of medical imaging data for research purposes.

Additional information

Correspondence and requests for materials should be addressed to M.F.