

Article

One-Pixel Attacks Can Improve the Correctness of Prediction

Wiktorja Tajak ¹, Adam Piórkowski ^{1,*} and Karolina Nurzyńska ²¹ Department of Biocybernetics and Biomedical Engineering, AGH University of Krakow, 30-059 Krakow, Poland² Algorithmics and Software Division, Silesian University of Technology, 44-100 Gliwice, Poland; knurzynska@polsl.pl

* Correspondence: pioro@agh.edu.pl

Abstract

Convolutional neural networks (CNNs) are widely used in medical image classification, yet their robustness to localized perturbations remains limited. This study evaluates one-pixel attacks on VGG16, MobileNetV2, and EfficientNetV2-B0 using brain tumor MRI images resized to 96×96 pixels. Each pixel was systematically perturbed across grayscale intensities, and model responses were analyzed in terms of vulnerability, recoverability, and pixel-level sensitivity. The relationship between prediction confidence and influential pixel locations was also examined. Results show that all models remain vulnerable to one-pixel perturbations despite high accuracy. Misclassified samples exhibit more successful attack locations, while correctly classified samples are more robust. Higher-intensity perturbations more often restore correct predictions in misclassified cases. A monotonic relationship is observed between prediction confidence and pixel sensitivity, where lower confidence corresponds to more influential pixels. Recovery points show spatially concentrated patterns. Overall, pixel-level sensitivity is more strongly associated with prediction correctness and local perturbations than with confidence. These findings are consistent across architectures and suggest that one-pixel analysis is useful for assessing CNN robustness in medical imaging.

Keywords: one-pixel attacks; adversarial attacks; artificial intelligence; convolutional neural networks; medical imaging

1. Introduction

Recent advances in deep learning have led to widespread adoption of convolutional neural networks (CNNs) for automated analysis of clinical imaging data [1–6]. These models achieve state-of-the-art performance in tasks such as tumor classification, segmentation, and disease detection, often reaching or exceeding human-level accuracy [5,7–15]. As a result, they are increasingly integrated into safety-critical diagnostic workflows [7,16–19].

Despite these advances, the reliability of deep learning systems in real-world clinical deployment remains limited by their vulnerability to adversarial attacks [3,4,20–25], alongside other data and evaluation-related challenges reported in the literature [2,26,27]. These attacks involve carefully crafted, often imperceptible perturbations [11,18,28,29] that can reliably mislead deep neural networks (DNNs) into making high-confidence misclassifications [8,10,14,30]. In medical applications, such failures extend beyond technical

Academic Editor: Miguel Alcaraz Baños

Received: 8 June 2026

Revised: 10 July 2026

Accepted: 10 July 2026

Published: 12 July 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](#) license.

errors and may lead to misdiagnosis, inappropriate treatment decisions, and other clinically significant risks [2,6,11,14,31].

A particularly stealthy and efficient class of such threats is the one-pixel attack (OPA) [18,25,31–33]. OPA is typically formulated as a black-box optimization problem that requires minimal knowledge of the target model architecture [2,11–13]. The attack identifies a single pixel whose modification maximizes misclassification probability, commonly using differential evolution (DE) [3,9,10,14,30,34], although random and heuristic search strategies have also been reported [1,9–11,18,20,32]. This class of attacks is especially relevant for grayscale imaging modalities such as magnetic resonance imaging (MRI), where isolated pixel perturbations may be indistinguishable from acquisition noise or imaging artifacts [1,15,19].

Current literature offers several competing explanations for the effectiveness of minimal perturbations. One view attributes model susceptibility to quasi-linear behavior in high-dimensional spaces [9,35,36], while alternative approaches emphasize the role of receptive fields and their sensitivity to localized disturbances in salient regions [1,13,30,31]. Additional studies propose “semantic decoupling” effects, in which adversarial perturbations shift model attention away from diagnostically relevant structures toward spurious background features [7]. It has also been suggested that medical imaging models may exhibit increased vulnerability due to domain-specific texture properties and highly non-convex optimization landscapes associated with over-parameterized architectures [4,5,31].

Building on these findings, previous work provided an initial baseline characterization of OPA sensitivity in brain tumor MRI across multiple architectures, identifying MobileNetV2 and EfficientNetV2-B0 as being more robust than traditional architectures like VGG16 [1]. That study highlighted fragmented decision boundaries and reliance on non-diagnostic features. However, the analysis was limited to fixed-intensity perturbations, primarily focusing on maximum-intensity (“white pixel”) modifications, leaving the role of intensity variation insufficiently explored.

The present study extends this direction by systematically evaluating one-pixel perturbations across the full grayscale intensity range [0–255]. The analysis is conducted on VGG16, MobileNetV2, and EfficientNetV2-B0, selected based on previously reported robustness. The objective is to determine whether model vulnerability depends on perturbation intensity and to assess whether any architecture maintains stable performance across the full intensity spectrum.

2. Materials and Methods

This study extends previous work [1] on a one-pixel attack vulnerability in brain tumor MRI classification by evaluating perturbations across the full grayscale intensity range [0–255] to determine whether model sensitivity depends on perturbation intensity. The experiments are conducted on VGG16 [37], MobileNetV2 [38], and EfficientNetV2-B0 [39], using a consistent preprocessing and evaluation pipeline to compare robustness across architectures under continuous intensity changes. A more detailed description of the employed models can be found in [1].

2.1. Dataset

In continuation of our previous study [1], the same two publicly available brain tumor MRI datasets from Kaggle [40,41] were used. Before merging, duplicate images were removed using pixel-wise comparison, where images with identical pixel intensity matrices were detected and discarded. After this step, both datasets were merged into a single combined dataset for analysis. Table 1 presents the resulting merged dataset, including class distributions and the number of images in the training, validation, and testing sets

following a 7:2:1 stratified split with respect to class labels. Figure 1 shows representative examples of MRI scans in the axial, coronal, and sagittal planes. As in the original sources, the data consist of T1-weighted, T2-weighted, T1-weighted contrast-enhanced, and FLAIR sequences representing glioma, meningioma, pituitary, and normal (no tumor) classes. Following deduplication, the image resolutions ranged from 150×167 to 512×512 pixels.

Table 1. Dataset overview.

Class	Training	Validation	Testing
Glioma	1618	478	216
Meningioma	1072	302	157
Normal	1194	320	192
Pituitary	1235	369	160
Total	5119	1469	725

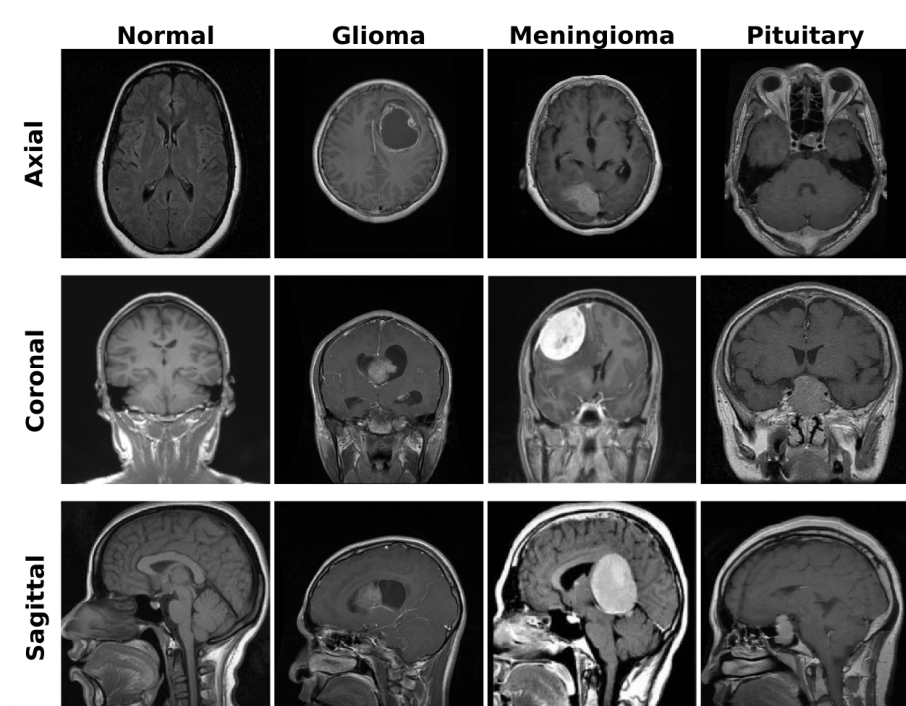


Figure 1. Exemplary images from the dataset.

2.2. Data Preprocessing

All MRI images were resized to a uniform resolution of 96×96 pixels to ensure consistent input dimensions across all models. Pixel intensity values were then normalized to the range $[0-1]$ to stabilize and accelerate the training process. Since the dataset consisted of grayscale MRI scans, no additional color space conversion was required. This preprocessing pipeline was applied uniformly across the entire merged dataset prior to training.

2.3. Training Setup and Model Configuration

The models were implemented using the Keras API with TensorFlow 2.10.0 as the backend. Transfer learning was applied by initializing the networks with pretrained ImageNet weights, which were then fine-tuned for the target task. The final classification head was adapted from the original architecture designs to support four output classes and consisted of a model-dependent feature aggregation layer (global average pooling or flatten), followed by one or more fully connected layers with ReLU activation. Dropout regularization was applied where appropriate to reduce overfitting, and the network was concluded with a softmax layer producing probability distributions over the four classes.

All models were trained using the Adam optimizer with categorical cross-entropy as the loss function. Performance was evaluated using accuracy, precision, recall, and F1-score. Training was performed with a batch size of 64 for up to 100 epochs, with early stopping triggered if validation loss did not improve for 5 consecutive epochs. Additionally, a learning rate scheduling strategy was applied, starting at 1×10^{-3} and reducing progressively down to a minimum of $1e-6$ when validation loss plateaued for 2 epochs.

Experiments were carried out in a Jupyter Notebook 7.3.2 environment using Visual Studio Code 1.128.0, Python 3.10.13, and the Anaconda distribution. Computations were executed on a workstation equipped with an Intel® Core™ i7-9700KF CPU (3.60 GHz), NVIDIA GeForce RTX 2070 SUPER (8 GB), and 64 GB of RAM.

2.4. Training Procedure and Evaluation Metrics

Each model was trained in three independent runs using the same training configuration. Model performance was evaluated consistently across all runs, and results were reported as mean $\pm$ standard deviation to account for run-to-run variability.

Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. Accuracy reflected the overall proportion of correctly classified samples, while precision measured the reliability of positive predictions. Recall quantified the model's ability to correctly identify all relevant cases, and the F1-score provided a balanced measure combining precision and recall. The confusion matrix was additionally used to analyze class-wise prediction performance and misclassification patterns across tumor categories. Class-specific precision, recall, and F1-score values were computed by treating each tumor category as a separate binary classification problem against all remaining classes.

2.5. Robustness Evaluation Under One-Pixel Attacks

A single trained model instance was selected as a fixed baseline for adversarial robustness evaluation. This representative model was chosen post hoc based on a combined performance score defined as the average of validation accuracy and macro-averaged F1-score. This procedure ensured a stable and representative model instance for computationally intensive adversarial experiments, without affecting the statistical reporting of classification performance across all runs.

To assess model robustness, a one-pixel attack was applied as an adversarial evaluation method. Unlike standard implementations limited to binary intensity changes, the attack was extended to the full grayscale range, enabling a fine-grained analysis of pixel-level sensitivity. Since the input images were normalized to the range $[0, 1]$ prior to model processing, all adversarial pixel perturbations defined in the $[0-255]$ space were appropriately rescaled to match the normalized input domain before being applied to the images. The overall procedure of the attack is illustrated in Figure 2.

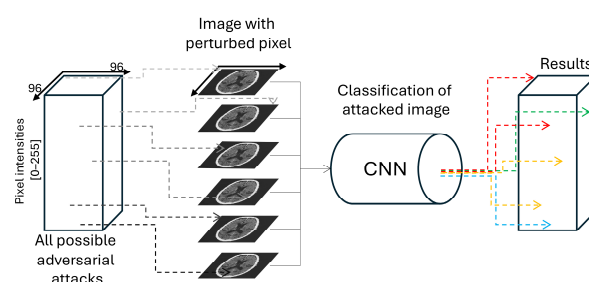


Figure 2. Overview of the one-pixel attack procedure. Each pixel is systematically perturbed across grayscale values $[0-255]$, generating 256 model response maps per image, which are stored as 2D outputs and a 3D intensity volume.

For each classified image, regardless of whether the original classification was correct or incorrect, every pixel location was systematically perturbed by assigning all possible grayscale intensity values while keeping the remaining pixels unchanged. Each modified image was then passed through the trained model, and predictions were recorded. This process resulted in 256 response maps per image, capturing the model's sensitivity to localized intensity variations across the entire grayscale range. The resulting maps were exported as PNG files for 2D visualization and as a TIFF (tagged image file format) stack representing a 3D intensity volume spanning all grayscale values, enabling subsequent volumetric analysis and 3D visualization.

Model robustness under attack was quantified using two complementary metrics: *vulnerability* and *recoverability*.

Vulnerability (V) was defined as a general measure of sensitivity to one-pixel perturbations, computed for both correctly and incorrectly classified samples. It quantifies whether at least one pixel-level perturbation is sufficient to change the original model prediction:

$$V = \frac{N_{vulnerable}}{N_{total}} \cdot 100\%, \quad (1)$$

where $N_{vulnerable}$ denotes the number of images for which at least one one-pixel perturbation results in a change of the original prediction, and N_{total} is the total number of evaluated images.

For correctly classified samples, vulnerability corresponds to cases where a perturbation induces misclassification, while for initially misclassified samples, it reflects cases where the prediction changes under perturbation (including potential correction).

Invulnerability was computed as the complement of vulnerability ($100\% - V$) for descriptive purposes only and was not used as an independent evaluation metric.

Recoverability (R) was defined specifically for initially vulnerable misclassified samples and measures the proportion of such samples for which at least one one-pixel perturbation restores the correct label:

$$R = \frac{N_{corrected}}{N_{vulnerable_incorrect}} \cdot 100\%, \quad (2)$$

where $N_{corrected}$ denotes the number of misclassified images whose prediction was corrected by at least one perturbation, and $N_{vulnerable_incorrect}$ is the total number of initially misclassified vulnerable samples.

3. Results

This section reports the experimental findings, encompassing training performance, robustness under one-pixel attack, and a detailed analysis of pixel-level sensitivity across the full intensity range.

3.1. Training Performance and Evaluation Metrics

Table 2 summarizes the performance of the evaluated models across training, validation, and test settings, including class-wise precision, recall, and F1-score. MobileNetV2 achieved the highest overall performance, with a test accuracy of 0.946 ± 0.007 , closely followed by EfficientNetV2-B0, which reached 0.937 ± 0.004 . VGG16 showed comparatively lower performance, with a test accuracy of 0.902 ± 0.005 .

Table 2. Training results and evaluation metrics (given as mean $\pm$ standard deviation).

Measure	Model	VGG16	MobileNetV2	EfficientNetV2-B0
Training accuracy		0.917 ± 0.003	0.984 ± 0.003	0.948 ± 0.008
Validation accuracy		0.904 ± 0.001	0.946 ± 0.004	0.932 ± 0.005
Test accuracy		0.902 ± 0.005	0.946 ± 0.007	0.937 ± 0.004

Precision	Normal	0.950 ± 0.000	0.987 ± 0.005	0.987 ± 0.005
	Glioma	0.880 ± 0.008	0.943 ± 0.005	0.917 ± 0.009
	Meningioma	0.823 ± 0.005	0.903 ± 0.012	0.880 ± 0.000
	Pituitary	0.940 ± 0.008	0.943 ± 0.009	0.947 ± 0.005
Recall	Normal	0.990 ± 0.000	0.977 ± 0.005	0.983 ± 0.005
	Glioma	0.860 ± 0.000	0.933 ± 0.009	0.927 ± 0.005
	Meningioma	0.777 ± 0.021	0.873 ± 0.017	0.833 ± 0.012
	Pituitary	0.973 ± 0.005	0.990 ± 0.000	0.990 ± 0.000
F1-Score	Normal	0.970 ± 0.000	0.980 ± 0.000	0.987 ± 0.005
	Glioma	0.867 ± 0.005	0.940 ± 0.008	0.920 ± 0.008
	Meningioma	0.800 ± 0.016	0.887 ± 0.009	0.853 ± 0.005
	Pituitary	0.957 ± 0.005	0.967 ± 0.005	0.967 ± 0.005

Across models, consistent trends were observed between evaluation metrics, with MobileNetV2 and EfficientNetV2-B0 maintaining high precision and recall across all classes. For instance, both architectures achieved high precision for the normal class (0.987 ± 0.005), while also preserving competitive recall for challenging categories such as meningioma (0.903 ± 0.012 for MobileNetV2 and 0.880 ± 0.000 for EfficientNetV2-B0). In contrast, VGG16 exhibited weaker class-wise performance, particularly for meningioma, where both precision (0.823 ± 0.005) and recall (0.777 ± 0.021) were noticeably lower.

All reported results represent mean values and standard deviations computed over three independent training runs.

Overall, the results indicate that more recent architectures provide improved generalization and more stable performance across tumor classes, with EfficientNetV2-B0 and MobileNetV2 demonstrating the most consistent behavior across all evaluated metrics.

3.2. Vulnerability and Recoverability Analysis

Tables 3–5 and Figure 3 summarize the vulnerability and recoverability of the evaluated architectures under one-pixel attacks. A clear distinction can be observed between correctly and incorrectly classified samples. For correctly classified images, vulnerability remained relatively low for the normal ($\sim 1\text{--}2\%$), glioma ($\sim 3\text{--}10\%$), and pituitary ($\sim 2\text{--}10\%$) classes, whereas meningioma consistently exhibited the highest susceptibility to perturbations across all models ($\sim 29\text{--}37\%$). This trend was particularly pronounced for VGG16 and MobileNetV2, where approximately one-third of correctly classified meningioma images could be altered by a single-pixel modification.

Table 3. Vulnerability and recoverability for VGG16.

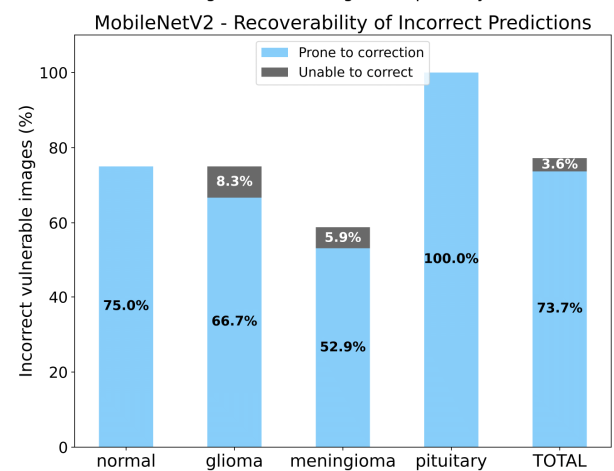
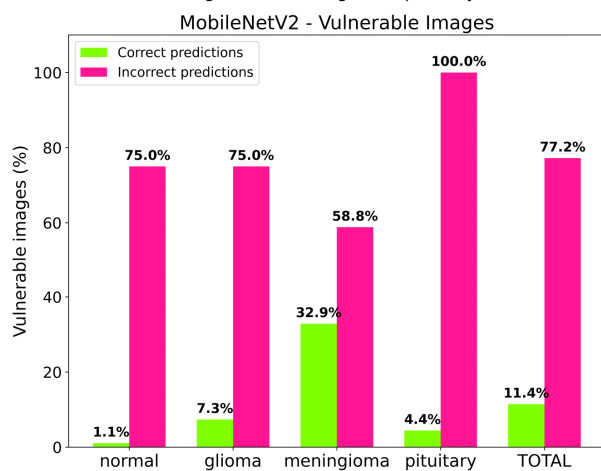
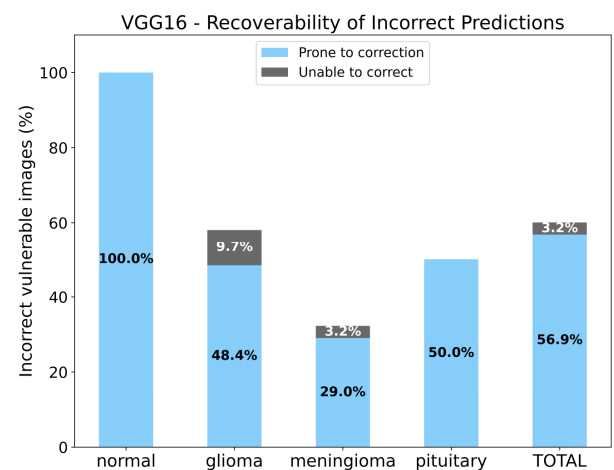
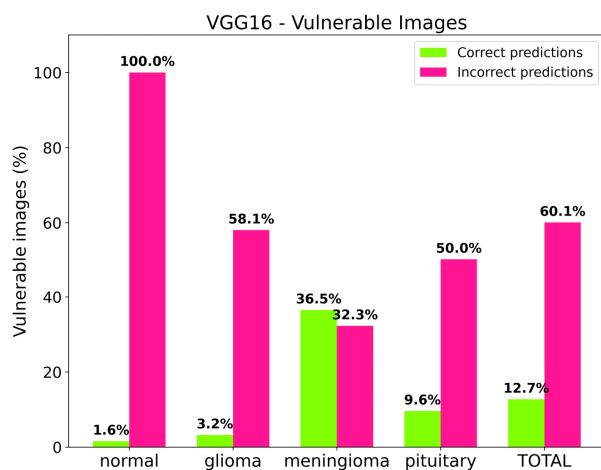
VGG16							
Prediction Type		Vulnerable Images	Vulnerability [%]	Invulnerable Images	Invulnerability [%]	Recoverable Images	Recoverability [%]
Correct	Normal	3	1.57	188	98.43	-	-
	Glioma	6	3.24	179	96.76	-	-
	Meningioma	46	36.51	80	63.49	-	-
	Pituitary	15	9.62	141	90.38	-	-
Incorrect	Normal	1	100.00	0	0.00	1	100.00
	Glioma	18	58.06	13	41.94	15	48.39
	Meningioma	10	32.26	21	67.74	9	29.03
	Pituitary	2	50.00	2	50.00	2	50.00

Table 4. Vulnerability and recoverability for MobileNetV2.

MobileNetV2							
Prediction Type		Vulnerable Images	Vulnerability [%]	Invulnerable Images	Invulnerability [%]	Recoverable Images	Recoverability [%]
Correct	Normal	2	1.58	186	98.42	-	-
	Glioma	15	7.35	189	92.65	-	-
	Meningioma	46	32.86	94	67.14	-	-
	Pituitary	7	4.40	152	95.60	-	-
Incorrect	Normal	3	75.00	1	25.00	3	75.00
	Glioma	9	75.00	3	25.00	8	66.67
	Meningioma	10	58.82	7	41.18	9	52.94
	Pituitary	1	100.00	0	0.00	1	100.00

Table 5. Vulnerability and recoverability for EfficientNetV2-B0.

EfficientNetV2-B0							
Prediction Type		Vulnerable Images	Vulnerability [%]	Invulnerable Images	Invulnerability [%]	Recoverable Images	Recoverability [%]
Correct	Normal	3	1.58	187	98.42	-	-
	Glioma	21	10.50	179	89.50	-	-
	Meningioma	39	29.10	95	70.90	-	-
	Pituitary	4	2.52	155	97.48	-	-
Incorrect	Normal	2	100.00	0	0.00	2	100.00
	Glioma	11	68.75	5	31.25	10	62.50
	Meningioma	14	60.87	9	39.13	13	56.52
	Pituitary	1	100.00	0	0.00	1	100.00



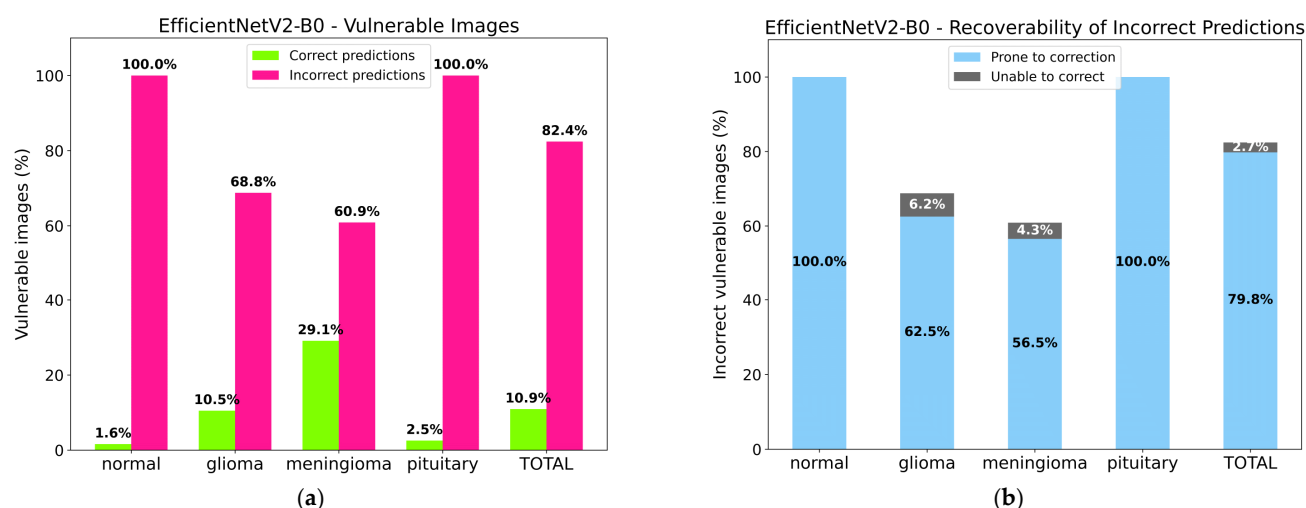


Figure 3. Image vulnerability and recoverability across models (VGG16, MobileNetV2, EfficientNetV2-B0): (a) Vulnerability and classification outcomes (correct vs. incorrect predictions) and (b) recoverability of incorrect predictions.

In contrast, incorrectly classified images were substantially more vulnerable to perturbations. Across all architectures, vulnerability frequently exceeded 50% and reached 100% for several class-model combinations. This suggests that samples located near decision boundaries are considerably more sensitive to localized intensity changes than correctly classified cases.

The recoverability analysis presented in Figure 3b reveals that a large proportion of incorrect predictions could be corrected through pixel-level perturbations. MobileNetV2 and EfficientNetV2-B0 demonstrated the highest overall recoverability, correcting approximately 74% and 80% of vulnerable misclassified images, respectively, compared with 57% for VGG16. Complete recoverability (100%) was observed for several class-specific cases, particularly within the normal and pituitary categories; however, this result was partly influenced by the relatively small number of misclassified samples in these groups.

Overall, the results indicate that one-pixel perturbations can both degrade and improve classification outcomes. While vulnerability varied substantially between tumor classes, recoverability remained consistently high for most vulnerable misclassified samples, highlighting the sensitivity of the learned decision boundaries to localized intensity modifications.

3.3. Pixel Intensity Analysis

A systematic pixel intensity analysis was conducted to evaluate the relationship between one-pixel adversarial perturbations and prediction recovery across different CNN architectures, including VGG16, MobileNetV2, and EfficientNetV2-B0. The full set of results is provided in Appendix A, while representative cases are shown in Figures 4 and 5.

As correctly classified images were largely insensitive to perturbation intensity, the analysis focuses on misclassified samples. Across all models and classes, increasing pixel intensity in one-pixel attacks consistently increased the probability of recovering the correct class prediction, indicating a generalizable pattern rather than model-specific behavior.

For VGG16 (Figure 4a), where the true class is pituitary, higher-intensity pixel perturbations increasingly result in correction of the prediction toward the true class. Simultaneously, the frequency of misclassification into the normal class decreases with rising intensity, while transitions involving glioma and meningioma remain largely unchanged.

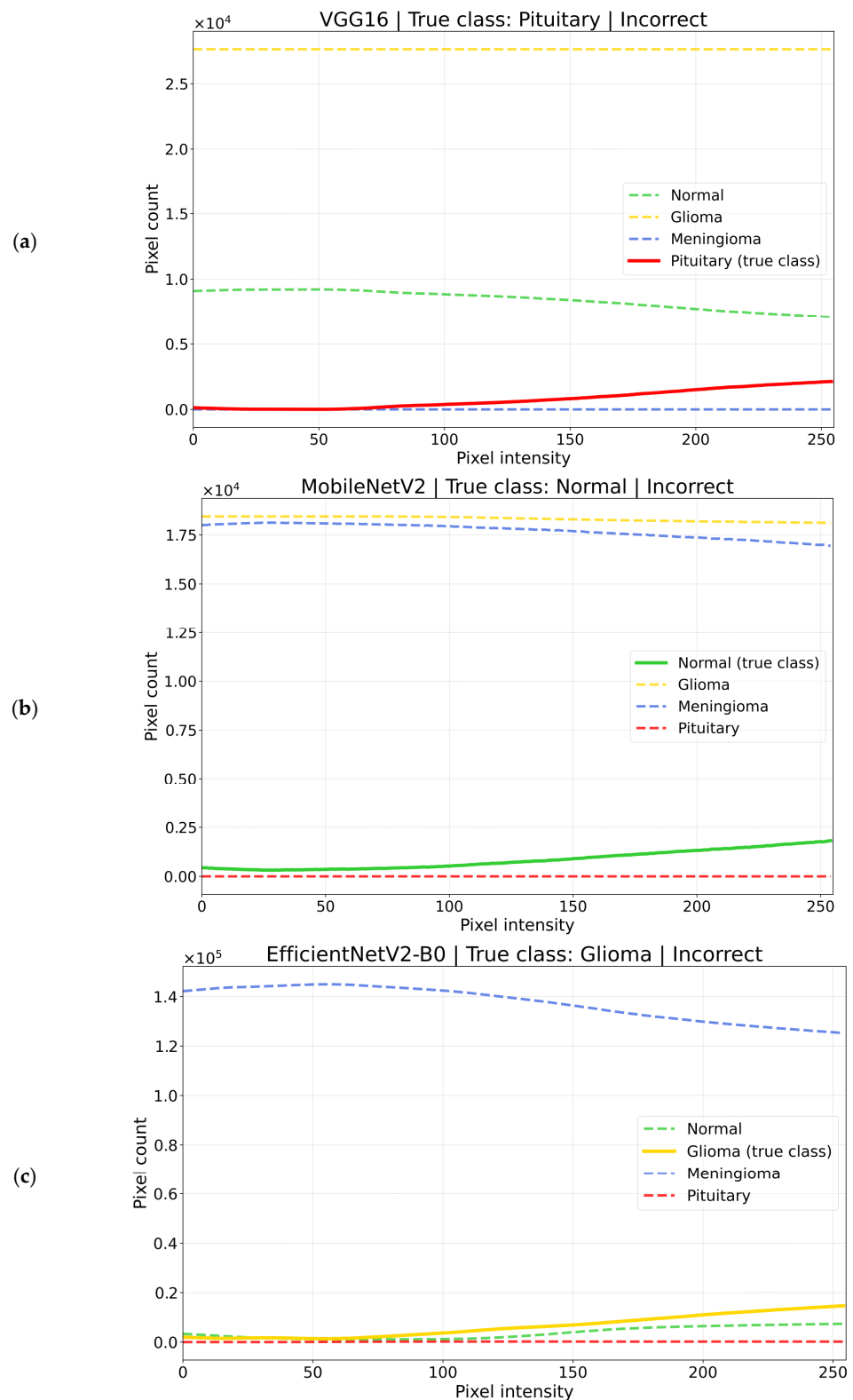


Figure 4. Pixel-intensity-dependent prediction recovery for incorrectly classified MRI images. The plots show the effect of increasing pixel intensity on transitions between predicted classes. The solid curve denotes the true class of the image, while dashed curves represent alternative class transitions. Examples are shown for (a) VGG16 (Pituitary), (b) MobileNetV2 (Normal), and (c) EfficientNetV2-B0 (Glioma).

In the case of MobileNetV2 (Figure 4b), with normal as the true class, a similar monotonic relationship is observed: stronger pixel perturbations increase the likelihood of re-

covery to the correct class. In contrast, transitions toward glioma and meningioma decrease with intensity, with meningioma exhibiting the strongest decline. The pituitary trajectory remains stable across intensity levels.

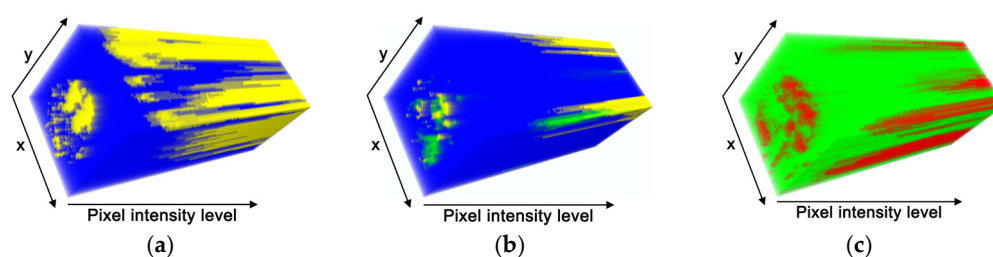


Figure 5. Geometric representation of pixel-space distributions and adversarial recovery via pixel perturbations, representing the most commonly observed empirical pattern in this study. Data blocks represent 96×96 pixel images embedded in a 3D space, where the third axis encodes pixel intensity (0 = front, 255 = back). (a) Glioma sample misclassified as meningioma, corrected to the true class via pixel-level perturbations (yellow recovery points); (b) another glioma sample misclassified as meningioma, either corrected to the true class (yellow) or driven to normal (green); (c) pituitary sample misclassified as normal, recovered to the true class via pixel perturbations (red).

A similar trend is also observed for EfficientNetV2-B0 (Figure 4c), where the true class is glioma. Increasing pixel intensity again enhances recoverability of the correct prediction. While a slight increase in secondary transitions to normal and pituitary is visible, misclassification into meningioma decreases markedly with increasing intensity.

Overall, the results indicate a strong and consistent relationship between pixel intensity and prediction recovery, suggesting that higher-intensity perturbations are more effective in correcting incorrect classifications across all evaluated architectures.

A more detailed volumetric interpretation of this phenomenon is provided in Figure 5. In the 3D visualization of the influence of the intensity change and location of the perturbed pixel, recovery points become increasingly dense at higher intensity levels. These points are not uniformly distributed across the image plane and show a tendency toward spatial concentration in peripheral or background areas, particularly at higher intensities. This observation is further supported by the aggregate spatial distributions presented in Figure 6. At lower intensity levels, more mixed distributions are observed, including both correct and incorrect prediction points within anatomically relevant regions.

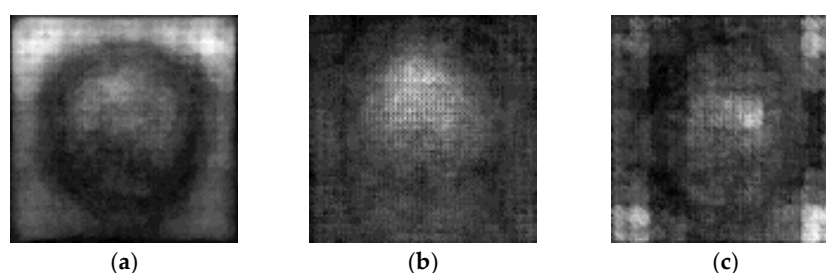


Figure 6. Statistical spatial maps of pixel attack occurrence in the 96×96 image plane, computed over all test images for (a) VGG16, (b) MobileNetV2, and (c) EfficientNetV2-B0.

Specifically, in Figure 5a, recovery points for glioma are sparse at low intensity levels and show increased concentration toward peripheral regions at higher intensities. Figure 5b exhibits a more uniform distribution of recovery points across both anatomical and non-anatomical regions. In Figure 5c, recovery points for the pituitary initially appear within the brain region at lower intensities and show a tendency toward peripheral regions at higher intensities.

Correctly classified samples show low sensitivity to pixel intensity perturbations and no systematic dependence on intensity level (see Appendix A for full results), confirming that the recoverability effect is specific to initially misclassified cases.

Finally, the selection of representative cases in Figures 4–5 reflects a balanced sampling across architectures and classes. While one case involved a single dominant misclassification instance, it nonetheless exhibited a strong manifestation of the described phenomenon, consistent with the broader dataset trends.

A statistical spatial aggregation of pixel attack occurrences across all test images is shown in Figure 6. The resulting heatmaps reveal architecture-dependent spatial distributions of attack-sensitive pixel locations. VGG16 (Figure 6a) exhibits a more diffuse pattern, with higher occurrence frequencies concentrated predominantly in peripheral and background regions. In contrast, MobileNetV2 (Figure 6b) shows a stronger concentration of attack-sensitive locations within anatomically relevant brain regions. EfficientNetV2-B0 (Figure 6c) presents an intermediate behavior, combining characteristics of both patterns with localized clusters observed in both central and peripheral areas. The intensity of each heatmap location reflects the number of images for which a successful attack was observed at the corresponding pixel coordinate, with brighter regions indicating higher occurrence frequencies. These results indicate that the spatial distribution of adversarially influential pixels is not uniform and varies across architectures, suggesting that spatial vulnerability patterns are influenced by the underlying network architecture.

Figure 7 presents the pixel intensity distributions extracted from adversarial attack maps generated from successful attacks on VGG16, MobileNetV2, and EfficientNetV2-B0. The histograms illustrate the minimal pixel-intensity differences associated with successful prediction changes for each architecture. Start-point markers indicate the lowest observed non-zero perturbation intensity present in successful adversarial attack maps. Across models, the minimal observed non-zero perturbation intensity is lowest for MobileNetV2 (2), followed by EfficientNetV2-B0 (3), and highest for VGG16 (8), indicating a model-dependent shift in the lower bound of observed successful adversarial perturbations. Differences in the shapes of these distributions suggest that successful attack patterns vary between models, indicating that adversarial perturbations manifest differently depending on the underlying network structure. This provides an empirical characterization of how attack-related pixel modifications are distributed rather than directly measuring the minimal perturbation required for misclassification.

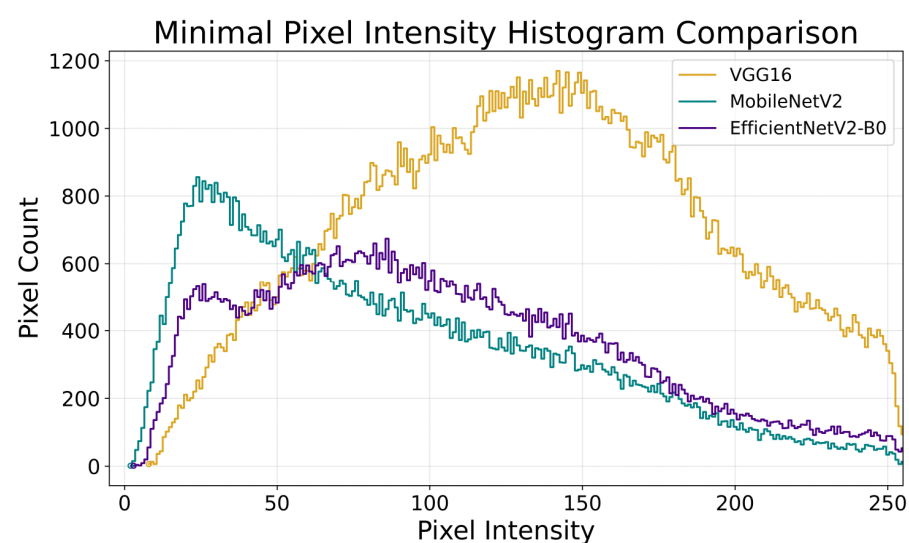


Figure 7. Minimal pixel intensity histograms (bin width = 1) derived from adversarial attack maps corresponding to successful attacks on VGG16, MobileNetV2, and EfficientNetV2-B0, showing the

distribution of pixel values associated with adversarial perturbations. Start-point markers indicate the minimal observed non-zero perturbation intensity in successful adversarial attack maps.

The influence of the intensity of the pixel triggering a successful attack should also be considered in relation to the semantic content of the image. To investigate this aspect, spatial maps were generated for each network using all 725 images from the test set. These maps represent the distribution of the minimum intensity change required to alter the model prediction for two scenarios: when the perturbing pixel was darker than the corresponding pixel in the original image, and when the perturbing pixel was brighter. For visualization purposes, a diverging colormap is used, where higher values are shown in red and lower values in blue. Black regions correspond to zero-valued entries.

The resulting maps, together with the minimum absolute intensity change derived from these values, are presented in Figure 8.

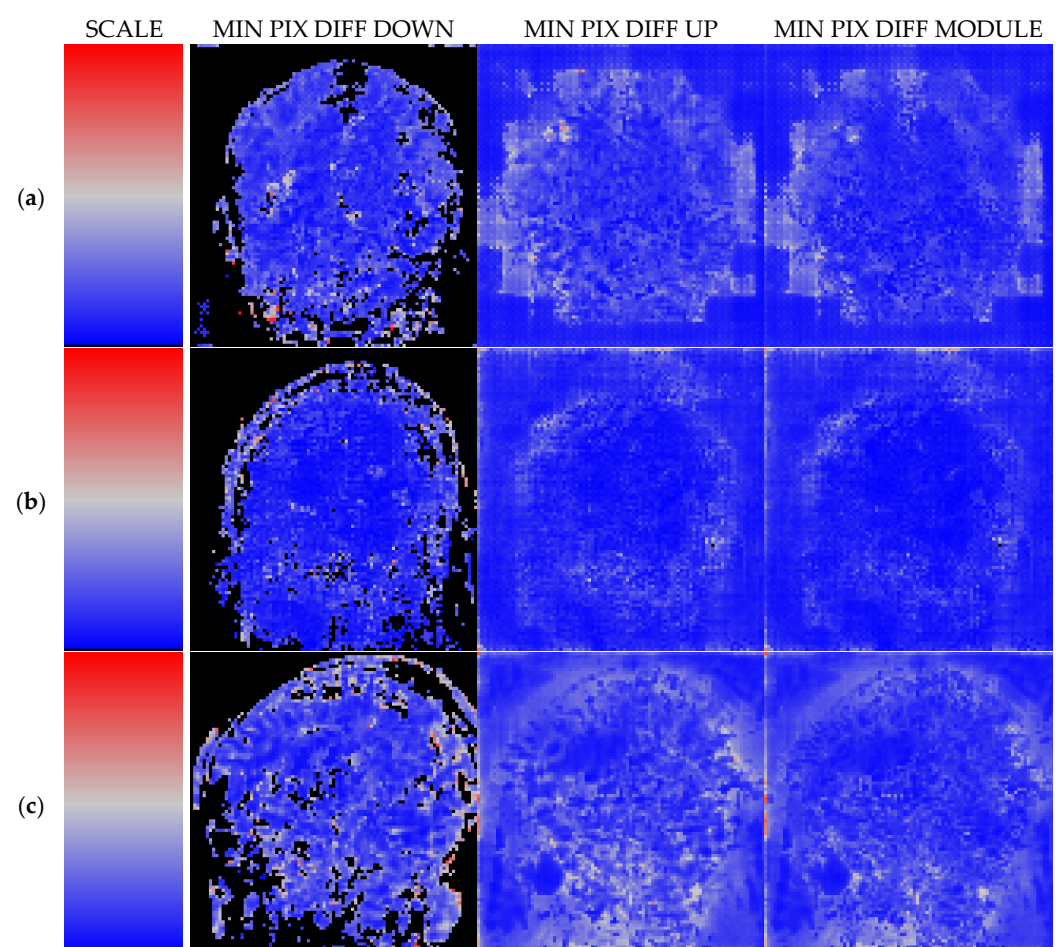


Figure 8. Spatial maps of the minimum effective pixel perturbation values aggregated across the entire test set for (a) VGG16, (b) MobileNetV2, and (c) EfficientNetV2-B0. The first column shows maps corresponding to attacks performed by decreasing the original pixel intensity (darkening), the second column shows attacks performed by increasing the original pixel intensity (brightening), and the third column presents the minimum absolute perturbation value obtained from the two preceding maps. Brighter regions indicate locations where larger perturbation magnitudes were required to successfully alter the model prediction. The visualization uses a diverging colormap, where red denotes higher values, blue denotes lower values, and black corresponds to zero-valued entries, indicating pixel locations that were never identified as successful attack sites across the test set.

Analysis of these maps indicates that attacks involving pixels darker than the original image are most frequently concentrated within the examined anatomical tissue. In contrast, attacks involving pixels brighter than the original image tend to be distributed over a broader spatial area, often extending beyond the patient's body region into the surrounding background.

It is also important to assess the extent to which one-pixel attacks affect the actual image content and to compare these effects with interpretation at the radiologist's level. Figure 9 presents representative examples for the three evaluated networks, showing images that were susceptible to attacks induced by both darker and brighter pixel perturbations.

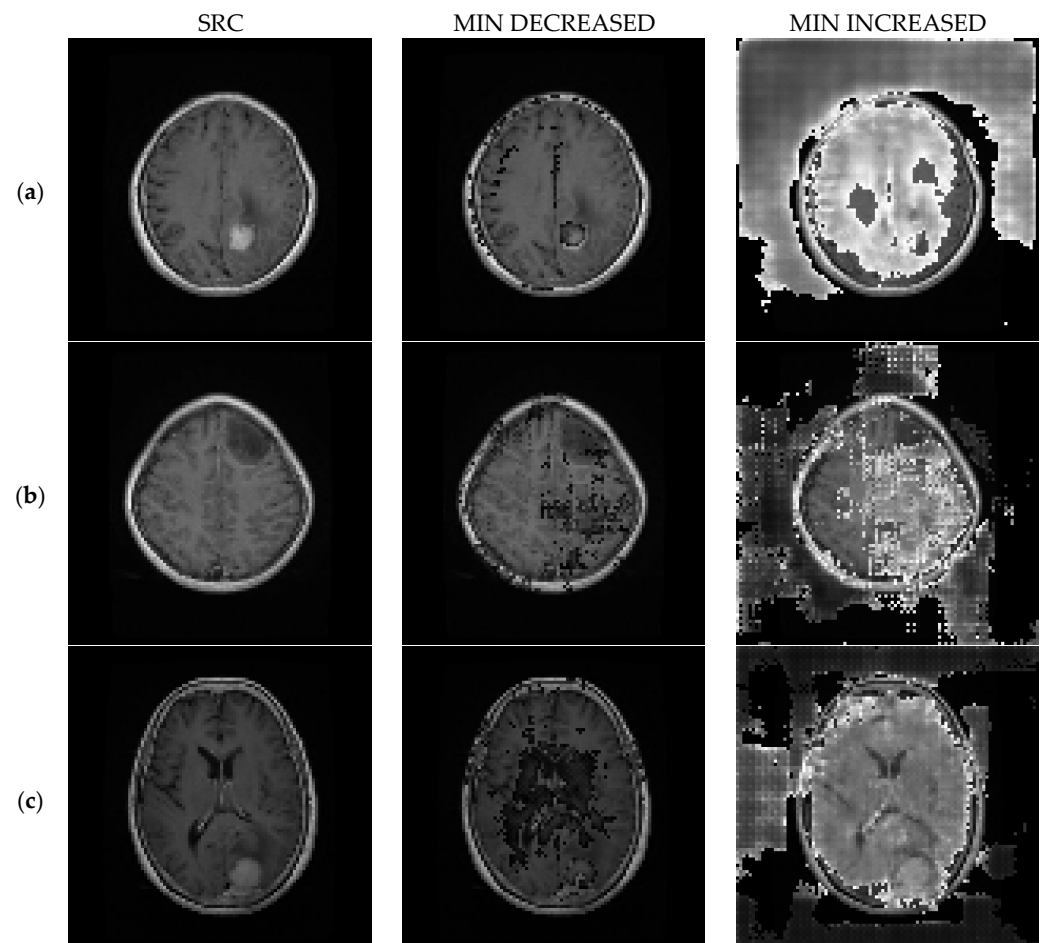


Figure 9. Examples of prediction changes induced by single-pixel perturbations. For each network, (a) VGG16, (b) MobileNetV2, and (c) EfficientNetV2-B0, the first column shows the original MRI image. The second column presents pixel locations where decreasing the original pixel intensity by the minimum effective amount resulted in a prediction change, while the third column shows pixel locations where increasing the original pixel intensity by the minimum effective amount altered the prediction. All highlighted locations correspond to successful one-pixel perturbations capable of changing the model's predicted class.

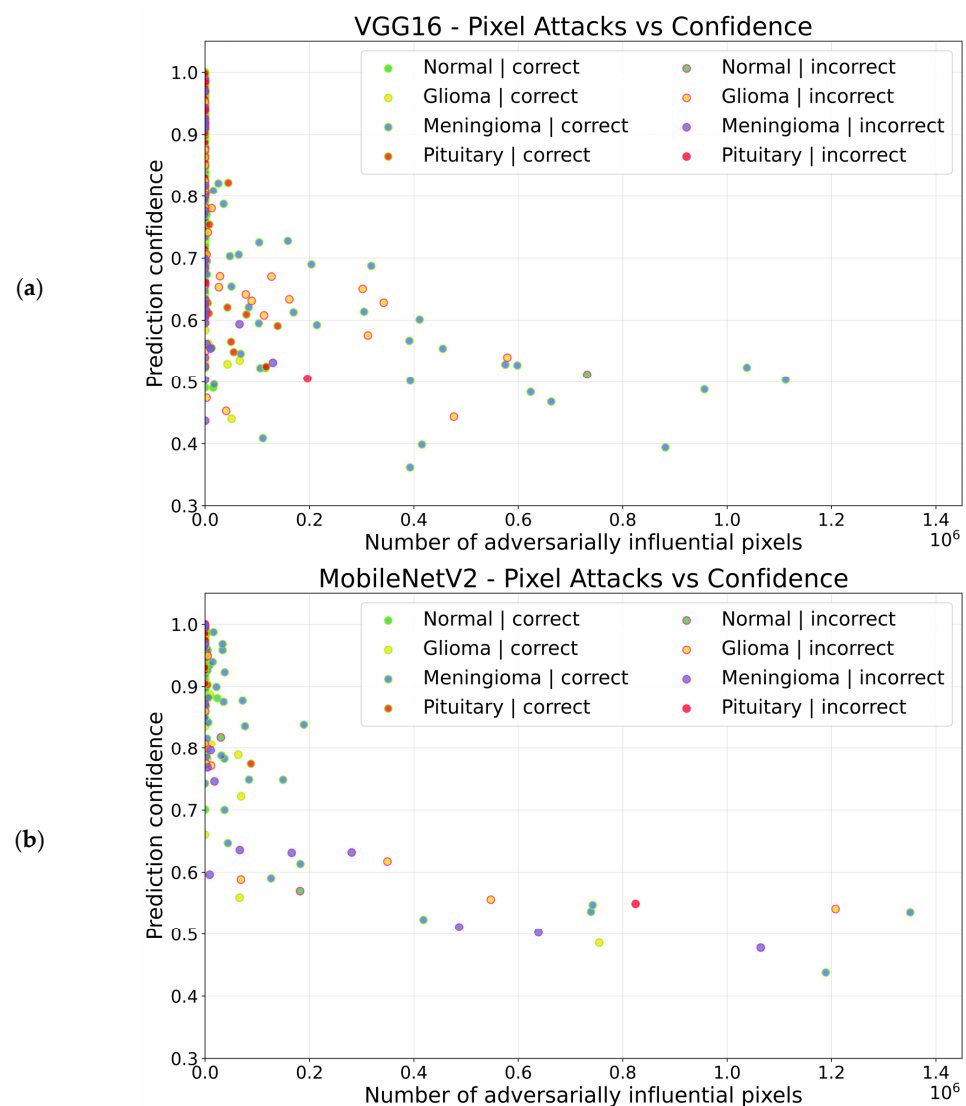
The first column contains the original MRI image, resized to 96×96 pixels for model input. In the second column, all pixel locations corresponding to the minimum effective perturbation values that successfully altered the prediction through a decrease in pixel intensity are overlaid on the original image. Similarly, the third column presents overlays of pixel locations corresponding to the minimum effective perturbation values required to alter the prediction through an increase in pixel intensity.

It should be emphasized that a successful one-pixel attack requires modifying only a single pixel relative to the original image. Nevertheless, the aggregated visualization of all effective locations provides insight into the spatial distribution of potential attack points. Based on these examples, it can be observed that prediction changes induced by pixel darkening tend to occur predominantly within the examined anatomical tissue, whereas prediction changes induced by pixel brightening are more frequently associated with locations outside the patient's body region.

3.4. Pixel Influence Versus Prediction Confidence

This subsection investigates the relationship between prediction confidence and the number of pixel locations enabling successful one-pixel attacks across VGG16, MobileNetV2, and EfficientNetV2-B0. The results are presented in Figure 10.

Across all architectures, a consistent monotonic trend is observed: lower prediction confidence is associated with a higher number of adversarially influential pixels, while higher confidence corresponds to a smaller set of pixel locations capable of altering the model's decision. Importantly, this relationship holds for both correctly and incorrectly classified samples. This indicates that even when the model is wrong, it can still assign high confidence to its (incorrect) prediction, which coincides with reduced sensitivity to single-pixel perturbations.



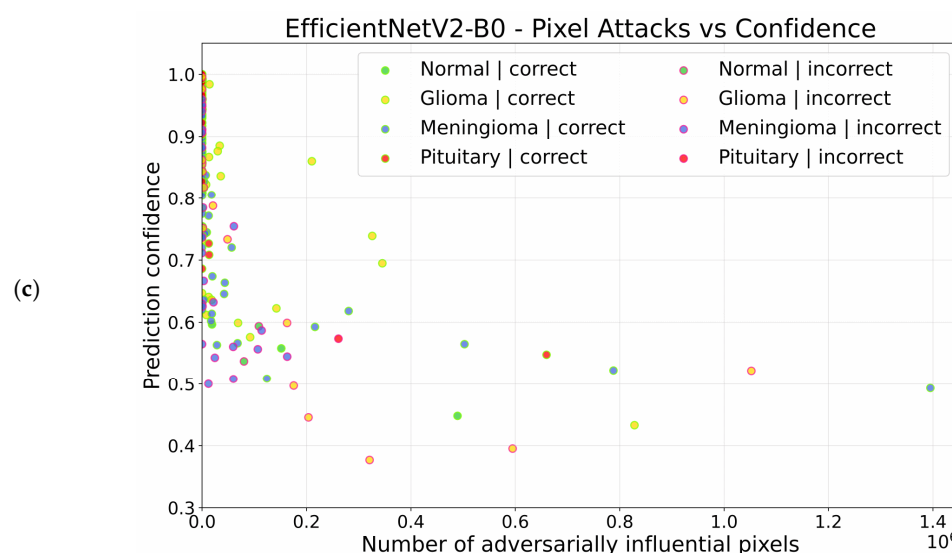


Figure 10. Pixel attack influence on prediction confidence across CNN architectures.

Relationship between number of influential adversarial pixels with different intensity (X-axis) and prediction confidence (Y-axis): (a) VGG16, (b) MobileNetV2, and (c) EfficientNetV2-B0.

The separation between correct and incorrect predictions is therefore weaker than initially assumed. Instead of forming distinct behavioral regimes, both groups follow a similar dependency structure, differing mainly in distribution and density rather than in trend direction.

Architecture-wise, VGG16 shows the widest dispersion of points, indicating higher variability in both confidence and susceptibility to pixel-level attacks. MobileNetV2 and EfficientNetV2-B0 exhibit more compact distributions, with clearer clustering at higher confidence and lower numbers of influential pixels, suggesting more stable behavior under perturbations.

Prior to analysis, normality was evaluated using the D'Agostino–Pearson test and Q–Q plots, indicating non-normal distributions for all variables; therefore, Spearman's rank correlation (two-sided, H_0 : no association) was applied. The strength of correlations was interpreted according to commonly used thresholds, where absolute correlation coefficients of 0.3–0.69 indicate moderate association and 0.7–0.99 indicate strong association. The results confirm a consistent negative relationship between prediction confidence and the number of adversarially influential pixels across all architectures (EfficientNetV2-B0: $\rho = -0.571$, $p < 1 \times 10^{-63}$; MobileNetV2: $\rho = -0.628$, $p < 1 \times 10^{-80}$; VGG16: $\rho = -0.569$, $p < 1 \times 10^{-63}$), corresponding to moderate negative correlations.

To further investigate whether this relationship differed depending on the original prediction outcome, separate correlations were calculated for correctly and incorrectly classified samples. For correctly classified samples, moderate negative correlations were observed across all architectures (EfficientNetV2-B0: $\rho = -0.501$, $p < 1 \times 10^{-44}$; MobileNetV2: $\rho = -0.572$, $p < 1 \times 10^{-61}$; VGG16: $\rho = -0.513$, $p < 1 \times 10^{-45}$). For initially misclassified samples, stronger negative associations were observed (EfficientNetV2-B0: $\rho = -0.830$, $p < 1 \times 10^{-11}$; MobileNetV2: $\rho = -0.919$, $p < 1 \times 10^{-14}$; VGG16: $\rho = -0.617$, $p < 1 \times 10^{-8}$). These results suggest that samples requiring fewer influential pixel changes tend to exhibit higher prediction confidence, with the relationship being particularly pronounced among initially misclassified samples. However, the correlations for misclassified subsets should be interpreted with consideration of their smaller sample sizes ($n = 34$ – 67).

4. Discussion

The obtained results highlight clear differences in performance, robustness, and pixel-level sensitivity across the evaluated CNN architectures.

From a performance perspective, newer architectures such as MobileNetV2 and EfficientNetV2-B0 consistently outperform VGG16 in terms of overall accuracy and class-wise balance. This improvement is particularly evident for more challenging categories such as meningioma, suggesting better feature extraction and generalization capabilities in modern architectures.

Despite higher classification performance, robustness to one-pixel adversarial attacks remains limited across all models. Misclassified samples are substantially more vulnerable than correctly classified ones, while recoverability remains high, indicating that decision boundaries are highly sensitive to localized perturbations.

A consistent relationship is observed between pixel intensity and recoverability. Across all architectures and classes, higher-intensity one-pixel perturbations increase the likelihood of correcting incorrect predictions. This suggests a structured sensitivity landscape, where perturbation strength plays a key role in determining prediction transitions.

More detailed insight emerges from the relationship between prediction confidence and pixel-level vulnerability. Correct predictions are associated with a low number of successful one-pixel attack locations, while incorrect predictions exhibit a substantially larger set of effective pixel positions. Importantly, this reflects the size of the effective attack space rather than the absolute ease of attack. The relationship between confidence and pixel-level sensitivity is monotonic rather than exhibiting any non-linear pattern.

Architecture-dependent differences are also evident. VGG16 shows a more dispersed distribution across both confidence and pixel sensitivity, reflecting higher variability in decision behavior. In contrast, MobileNetV2 and EfficientNetV2-B0 exhibit more compact and structured distributions, with clearer clustering of high-confidence correct predictions and more distinct separation between sensitive and insensitive regions.

A more detailed volumetric interpretation suggests a structured organization of recoverability within the 3D intensity space. Rather than being uniformly distributed, recovery points exhibit a spatial bias, with a tendency to concentrate in peripheral and background regions at higher intensity levels. This pattern is consistent across multiple cases, although with case-dependent variability, and is further supported by the aggregated spatial analysis presented in Figure 6.

In contrast, lower intensity levels show more dispersed behavior, with mixed distributions of correct and incorrect predictions across anatomically relevant regions, suggesting weaker spatial selectivity in this regime. The intensity-based distributions in Figure 7 further quantify these effects by illustrating model-dependent differences in perturbation intensity distributions associated with successful adversarial attacks.

The additional spatial analyses presented in Figures 8 and 9 further indicate that the direction of intensity modification influences the localization of effective perturbations. Pixel-darkening attacks are more frequently associated with anatomically relevant tissue regions, whereas pixel-brightening attacks tend to occur over a broader area, often extending into the image background. Although these observations were obtained from low-resolution images and should therefore be interpreted with caution, they suggest that pixel-level adversarial sensitivity is not entirely random and may be influenced by both image semantics and perturbation intensity.

These findings suggest that prediction correctness is more strongly associated with pixel-level adversarial sensitivity than prediction confidence. While confidence behavior varies across architectures, the underlying structure of one-pixel vulnerability remains consistent, suggesting that it is primarily governed by decision boundary geometry rather than model-specific noise.

4.1. Significance of the Study

This study provides new insights into the behavior of one-pixel adversarial attacks by demonstrating that minimal pixel-level perturbations can not only degrade but also correct misclassified predictions. This dual effect highlights a previously underexplored aspect of adversarial sensitivity, where single-pixel modifications reveal structured properties of decision boundaries rather than purely inducing failure.

The results further show a consistent relationship between prediction correctness and the number of successful one-pixel attack locations, suggesting that adversarial vulnerability is strongly linked to the geometry of the learned decision space. In addition, the observed relationship between prediction confidence and sensitivity is monotonic, indicating that confidence alone is insufficient to characterize model robustness.

Overall, the findings contribute to a deeper understanding of how CNN models respond to extreme localized perturbations and suggest that one-pixel analysis can serve as a useful tool for probing and interpreting model decision behavior.

4.2. Limitations of the Study

Several limitations should be acknowledged. A key limitation of this study is the use of publicly available Kaggle brain tumor MRI datasets, which do not provide detailed patient-level metadata (e.g., age, sex, or disease progression) and may therefore limit the clinical representativeness of the analyzed population. In addition, the fixed input resolution of 96×96 pixels constrains the spatial granularity of the analysis. At this scale, individual pixel perturbations may have a relatively strong impact, which can influence the observed sensitivity patterns and limit direct generalization to higher-resolution imaging scenarios. However, previous work [1], which evaluated seven CNN architectures on 224×224 inputs, showed that similar one-pixel sensitivity patterns are also present at higher resolutions, albeit with reduced intensity.

The present study also considers a limited subset of architectures and a single imaging domain. While consistent trends are observed across VGG16, MobileNetV2, and EfficientNetV2-B0, prior results [1] demonstrated comparable behavior across a broader set of seven CNN architectures under fixed single-pixel perturbations, suggesting that the observed phenomenon is not limited to a specific convolutional architecture. However, differences between CNN-based and Transformer-based models may influence sensitivity to localized pixel perturbations due to their distinct feature representation mechanisms [42,43]. Extending the analysis to Transformer-based architectures and additional datasets would provide further insight into whether the observed vulnerability and recoverability patterns are specific to convolutional models or represent a broader property of deep learning-based image classifiers.

Finally, while clear and consistent patterns are identified in terms of pixel-level sensitivity and recoverability, the underlying mechanisms remain open to further investigation, particularly in relation to decision boundary geometry and learned feature representations. The observed dependence on pixel intensity may be partly influenced by the inherent intensity distribution in MRI data, where higher intensities are more frequently associated with tissue regions, whereas lower intensities often correspond to background areas.

5. Conclusions

The main conclusions drawn from this study, conducted under a 96×96 image resolution and limited to the evaluated CNN architectures (MobileNetV2, EfficientNetV2-B0, and VGG16), are as follows:

Classification and Robustness Findings

- Despite high classification accuracy, all evaluated CNN models remain vulnerable to one-pixel adversarial attacks.
- Misclassified samples exhibit a higher number of successful one-pixel attack locations, indicating increased sensitivity to localized perturbations.
- Correctly classified samples are generally more robust, showing fewer successful attack locations.

Vulnerability and Recoverability Findings

- Pixel intensity affects prediction recovery, with higher-intensity perturbations more often restoring correct predictions.
- Lower prediction confidence corresponds to a higher number of influential pixel locations with different intensities.

Spatial Distribution and Sensitivity Patterns

- A relatively large proportion of influential pixel locations is observed outside the region corresponding to the examined anatomical structure.

It is important to note that the reported spatial and intensity-based patterns are derived from aggregated empirical distributions computed over the full test set (Figures 6–9), rather than from per-sample statistical hypothesis testing. As such, they represent population-level tendencies of adversarial sensitivity rather than deterministic spatial rules. These findings are consistent across all evaluated architectures and input subsets.

Overall, adversarial sensitivity is more strongly related to decision boundary structure than to prediction confidence, and this behavior is consistent across all tested architectures.

Future work should extend the analysis to additional CNN and transformer-based architectures, as well as datasets with different characteristics and resolutions, to assess the generality of the observed findings. It should also investigate additional perturbation baselines, including random single-pixel noise, to distinguish the effects of targeted gray-scale modifications from stochastic perturbations and provide a more comprehensive evaluation of model sensitivity to localized pixel-level changes. Such analyses would be consistent with previous studies demonstrating that localized pixel-level variations and image acquisition-related perturbations may influence deep learning model predictions in medical imaging tasks [44,45].

It would also be valuable to investigate data augmentation strategies such as noise injection and image inversion. In particular, combining datasets with different lighting conditions (e.g., dark and bright images) together with their inverted counterparts could help evaluate whether increased data diversity improves robustness and reduces sensitivity to localized perturbations.

Author Contributions: Conceptualization, W.T., K.N., and A.P.; methodology, W.T. and A.P.; software, W.T.; validation, W.T. and K.N.; formal analysis, W.T.; investigation, W.T.; resources, W.T.; data curation, A.P.; writing—original draft preparation, W.T.; writing—review and editing, W.T. and K.N.; visualization, W.T.; supervision, A.P. and K.N.; project administration, A.P.; funding acquisition, A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Public datasets were used.

Conflicts of Interest: The authors declare no conflicts of interest.

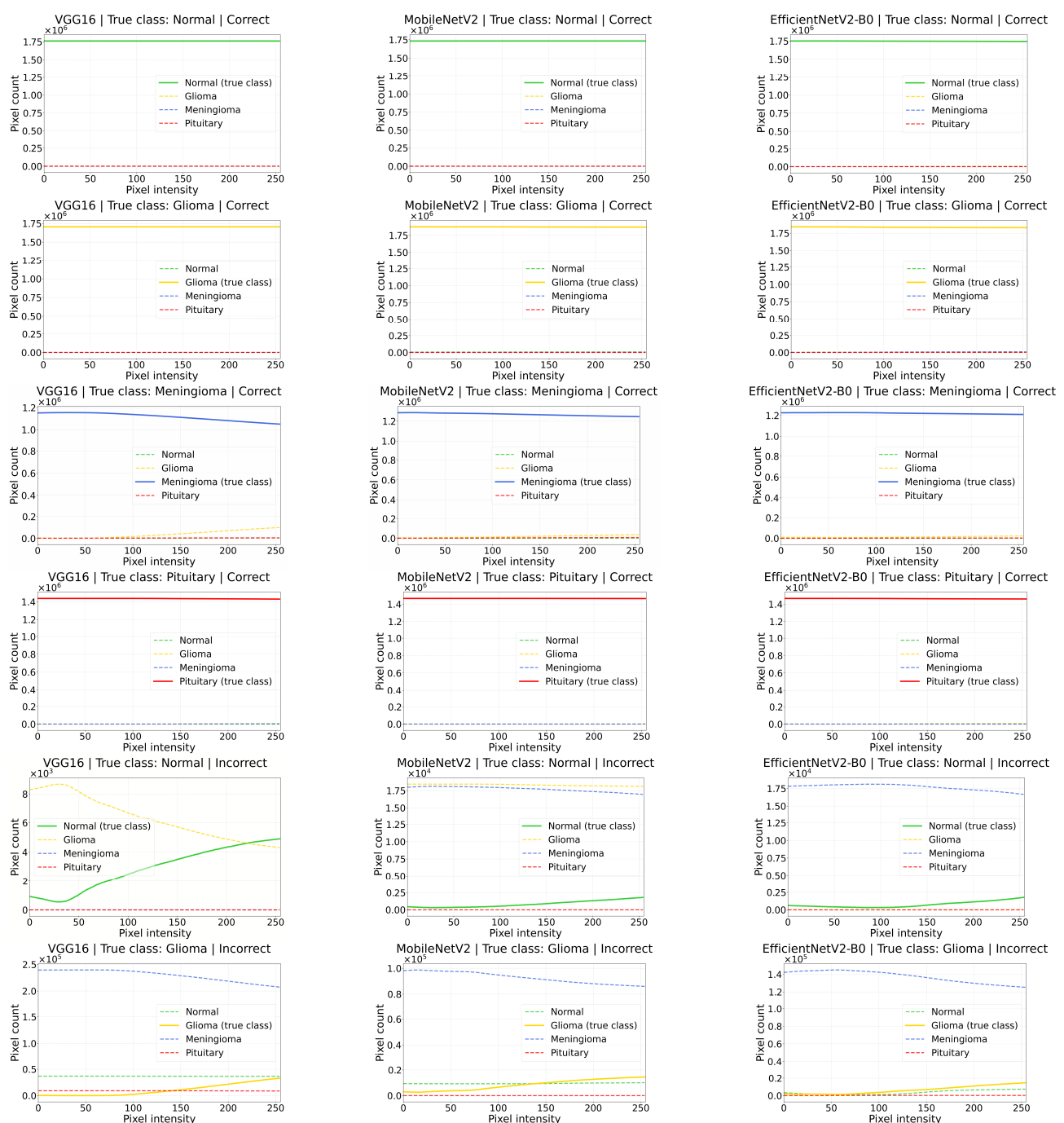
Abbreviations

The following abbreviations are used in this manuscript:

MRI	Magnetic Resonance Imaging
TIFF	Tagged Image File Format
OPA	One-Pixel Attack
CNN	Convolutional Neural Network
DNN	Deep Neural Network

Appendix A

All additional figures related to the pixel intensity analysis presented in Section 3.3 are included in this appendix. The complete set of visualizations illustrates the full range of pixel intensity levels and corresponding prediction transitions across all evaluated models and classes.



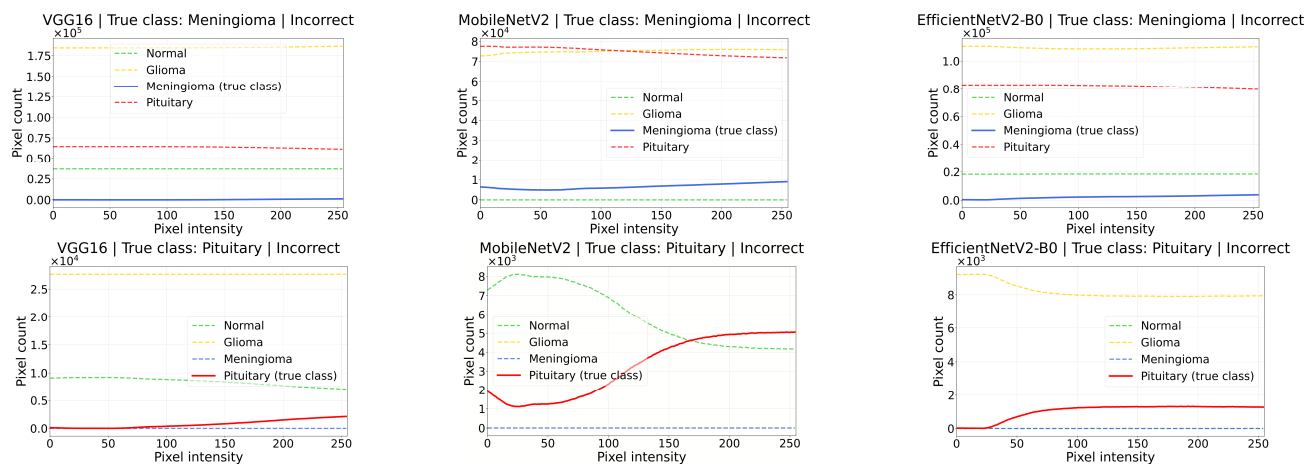


Figure A1. Complete collection of pixel-intensity-dependent prediction recovery trajectories for all evaluated CNN architectures and tumor classes. Each plot shows class-transition behavior under increasing pixel-intensity perturbations for initially misclassified MRI images. Solid lines correspond to the true class, whereas dashed lines indicate alternative predicted classes.

References

1. Tajak, W.; Nurzyńska, K.; Piórkowski, A. Vulnerability to One-Pixel Attacks of Neural Network Architectures in Medical Image Classification. *Bio-Algorithms Med-Syst.* **2025**, *21*, 58–70. <https://doi.org/10.5604/01.3001.0055.3261>.
2. Korpilahkola, J.; Sipola, T.; Kokkonen, T. Color-Optimized One-Pixel Attack Against Digital Pathology Images. In Proceedings of the 2021 29th Conference of Open Innovations Association (FRUCT), Tampere, Finland, 12–14 May 2021; pp. 206–213.
3. Husnoo, M.A.; Anwar, A. Do Not Get Fooled: Defense against the One-Pixel Attack to Protect IoT-Enabled Deep Learning Systems. *AD Hoc Netw.* **2021**, *122*, 102627. <https://doi.org/10.1016/j.adhoc.2021.102627>.
4. Hu, J.; He, Y.; Zhang, W.; Pang, S.; Ma, R.; Du, A. A Multimodal Adversarial Database: Towards A Comprehensive Assessment of Adversarial Attacks and Defenses on Medical Images. In Proceedings of the 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA), Thessaloniki, Greece, 9–13 October 2023; pp. 1–10.
5. Alzubaidi, L.; AL-Dulaimi, K.; Obeed, H.A.-H.; Saihood, A.; Fadhel, M.A.; Jebur, S.A.; Chen, Y.; Albahri, A.S.; Santamaría, J.; Gupta, A.; et al. MEFF—A Model Ensemble Feature Fusion Approach for Tackling Adversarial Attacks in Medical Imaging. *Intell. Syst. Appl.* **2024**, *22*, 200355. <https://doi.org/10.1016/j.iswa.2024.200355>.
6. Tsai, M.-J.; Lee, Y.-C.; Lien, H.-Y.; Liang, C.-C. Adversarial Defense for Medical Images. *Electronics* **2025**, *14*, 4384. <https://doi.org/10.3390/electronics14224384>.
7. Kamil, M.H.; Farma, E.P.T.; Basuki, S. Robustness Under Attack: Assessing Adversarial Fragility in Deep Learning Models for COVID-19 Radiography Prediction. *J. Electron. Electromed. Eng. Med. Inf.* **2026**, *8*, 769–788. <https://doi.org/10.35882/jeeemi.v8i2.1506>.
8. Quan, W.; Nagothu, D.; Poredi, N.A.; Chen, Y. CriPI: An Efficient Critical Pixels Identification Algorithm for Fast One-Pixel Attacks. In *Proceedings of the Sensors and Systems for Space Applications XIV*; Pham, K.D., Chen, G., Eds.; SPIE: Philadelphia, PA, USA, 12 April 2021; p. 21.
9. Wang, W.; Wildgruber, M.; Wang, Y. Attacking Medical Images with Minimal Noise: Exploiting Vulnerabilities in Medical Deep-Learning Systems. *Quant. Imaging Med. Surg.* **2024**, *14*, 9374–9384. <https://doi.org/10.21037/qims-24-1764>.
10. Nam, W.; Kim, K.; Moon, H.; Noh, H.; Park, J.; Kil, H. RISOPA: Rapid Imperceptible Strong One-Pixel Attacks in Deep Neural Networks. *Mathematics* **2024**, *12*, 1083. <https://doi.org/10.3390/math12071083>.
11. Paul, R.; Schabath, M.; Gillies, R.; Hall, L.; Goldgof, D. Mitigating Adversarial Attacks on Medical Image Understanding Systems. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 1517–1521.
12. Zhou, T.; Agrawal, S.; Manocha, P. Optimizing One-Pixel Black-Box Adversarial Attacks. *arXiv*, **2022**, arXiv:2205.02116.
13. Wang, Y.; Wang, W.; Wang, X.; Chen, Z. On One-Pixel Attacking Medical Images against Deep Learning Models. In Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science, Chengdu, China, 20 October 2023; pp. 248–257.

14. Tsai, M.-J.; Lin, P.-Y. Medical Images under Tampering. *Multimed. Tools Appl.* **2024**, *83*, 65407–65439. <https://doi.org/10.1007/s11042-023-17968-1>.
15. Huang, K.; Sahel, N.; Karki, D.; Xu, M.; Wang, Y. One Pixel Can Change the Diagnosis: Adversarial and Non-Adversarial Robustness and Uncertainty in Breast Ultrasound Classification Model. *Proc. AAAI Symp. Ser.* **2025**, *7*, 524–529.
16. Keppel, J.; Liebers, J.; Audal, J.; Gruenefeld, U.; Schneegass, S. ExplAinable Pixels: Investigating One-Pixel Attacks on Deep Learning Models with Explainable Visualizations. In Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia, Lisbon, Portugal, 27 November 2022; pp. 231–242.
17. Wang, P.; Li, Q.; Li, D.; Meng, S.; Bilal, M.; Mukherjee, A. Security in Defect Detection: A New One-Pixel Attack for Fooling DNNs. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 101689. <https://doi.org/10.1016/j.jksuci.2023.101689>.
18. Wang, Y.; Song, J.; Li, T.; Xin, Y.; Li, H.; Wei, N. Rectifying Multi-Attack Adversarial Perturbations in Deep Neural Network Based Image Classifier. *ACM Trans. Priv. Secur.* **2025**, *28*, 1–28. <https://doi.org/10.1145/3765757>.
19. Vali, E.; Alexandridis, G.; Stafylopatis, A. Adversarial Attacks & Detection on a Deep Learning-Based Digital Pathology Model. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
20. Alatalo, J.; Korpikalkola, J.; Sipola, T.; Kokkonen, T. Chromatic and Spatial Analysis of One-Pixel Attacks against an Image Classifier. In *Networked Systems. NETYS 2022*; Koulali, M.A., Mezini, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2022; Volume 13464, pp. 303–316.
21. Alatalo, J.; Sipola, T.; Kokkonen, T. Detecting One-Pixel Attacks Using Variational Autoencoders. In *Information Systems and Technologies*; Rocha, A., Adeli, H., Dzemyda, G., Moreira, F., Eds.; Lecture Notes in Networks and Systems; Springer International Publishing: Cham, Switzerland, 2022; Volume 468, pp. 611–623. ISBN 978-3-031-04825-8.
22. Khakurel, U.; Rawat, D.B. Adversarial Machine Learning Using Convolutional Neural Network With Imagenet. In Proceedings of the 2022 Annual Modeling and Simulation Conference (ANNSIM), San Diego, CA, USA, 18–20 July 2022; pp. 246–257.
23. Wei, H.; Tang, H.; Jia, X.; Wang, Z.; Yu, H.; Li, Z.; Satoh, S.; Van Gool, L.; Wang, Z. Physical Adversarial Attack Meets Computer Vision: A Decade Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 9797–9817. <https://doi.org/10.1109/TPAMI.2024.3430860>.
24. Zhu, J.; Zhao, H.; Yu, H.; Liu, J. Pixel Logo Attack: Embedding Attacks as Logo-Like Pixels. In Proceedings of the Genetic and Evolutionary Computation Conference, Melbourne, VIC, Australia, 14–18 July 2024; pp. 449–457.
25. Pal, S.; Yadav, G.; Jadidi, Z.; Habib, A.; Uddin, Md.P.; Karmakar, C.; Shukla, S. Vulnerabilities in Machine Learning for Cybersecurity: Current Trends and Future Research Directions. *J. Inf. Secur. Appl.* **2026**, *96*, 104269. <https://doi.org/10.1016/j.jisa.2025.104269>.
26. Klawikowska, Z.; Mikołajczyk, A.; Grochowski, M. Explainable AI for Inspecting Adversarial Attacks on Deep Neural Networks. In *Artificial Intelligence and Soft Computing*; Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J.M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 12415, pp. 134–146. ISBN 978-3-030-61400-3.
27. Leszczyńska, A.; Seweryn, M.; Obuchowicz, R.; Strzelecki, M.; Piórkowski, A.; Potocki, P.M. Artificial Intelligence for RECIST-Based Radiologic Treatment Response Assessment in Solid Tumors: A Systematic Review of Imaging- and Report-Derived Approaches. *Cancers* **2026**, *18*, 808. <https://doi.org/10.3390/cancers18050808>.
28. Li, C.; Wang, H.; Yao, W.; Jiang, T. Adversarial Attacks in Computer Vision: A Survey. *J. Membr. Comput.* **2024**, *6*, 130–147. <https://doi.org/10.1007/s41965-024-00142-3>.
29. Korpikalkola, J.; Sipola, T.; Puuska, S.; Kokkonen, T. One-Pixel Attack Deceives Computer-Assisted Diagnosis of Cancer. In Proceedings of the 2021 4th International Conference on Signal Processing and Machine Learning, Beijing, China, 18–20 August 2021; pp. 100–106.
30. Wang, W.; Sun, J.; Wang, G. Visualizing One Pixel Attack Using Adversarial Maps. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6 November 2020; pp. 924–929.
31. Tsai, M.-J.; Lin, P.-Y.; Lee, M.-E. Adversarial Attacks on Medical Image Classification. *Cancers* **2023**, *15*, 4228. <https://doi.org/10.3390/cancers15174228>.
32. Fan, H.; Qin, X.; Chen, S.; Shum, H.P.H.; Li, M. One-Index Vector Quantization Based Adversarial Attack on Image Classification. *Pattern Recognit. Lett.* **2024**, *186*, 47–56. <https://doi.org/10.1016/j.patrec.2024.09.001>.
33. Deng, S.; Zhou, X. One Pixel Is All I Need. *arXiv*, **2024**, arXiv:2412.10681.
34. Liu, Z.-Y.; Wang, P.S.; Hsiao, S.-C.; Tso, R. Defense against N-Pixel Attacks Based on Image Reconstruction. In Proceedings of the 8th International Workshop on Security in Blockchain and Cloud Computing, Taipei Taiwan, 6 October 2020; pp. 3–7.

35. Cucu, A.-V.; Valenzise, G.; Stănescu, D.; Ghergulescu, I.; Găină, L.I.; Gușiță, B. Defense Method against Adversarial Attacks Using JPEG Compression and One-Pixel Attack for Improved Dataset Security. In Proceedings of the 2023 27th International Conference on System Theory, Control and Computing (ICSTCC), Timisoara, Romania, 11–13 October 2023; pp. 523–527.
36. Kazmi, S.M.K.A.; Aafaq, N.; Khan, M.A.; Khalil, M.; Saleem, A. From Pixel to Peril: Investigating Adversarial Attacks on Aerial Imagery Through Comprehensive Review and Prospective Trajectories. *IEEE Access* **2023**, *11*, 81256–81278. <https://doi.org/10.1109/ACCESS.2023.3299878>.
37. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv*, **2015**, arXiv:1409.1556.
38. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv*, **2017**, arXiv:1704.04861.
39. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv*, **2020**, arXiv:1905.11946.
40. Nickparvar, M. Brain Tumor MRI Dataset. Available online: <https://www.kaggle.com/dsv/2645886> (accessed on 07 June 2026).
41. Bhuvaji, S.; Kadam, A.; Bhumkar, P.; Dedge, S.; Kanchan, S. Brain Tumor Classification (MRI). Available online: <https://www.kaggle.com/dsv/1183165> (accessed on 07 June 2026).
42. Aksoy, S.; Demircioglu, P.; Bogrekci, I. Systematic Failure of Vision Transformers in Imbalanced Skin Lesion Classification. *Dermato* **2026**, *6*, 22. <https://doi.org/10.3390/dermato6020022>.
43. He, K.; Gan, C.; Li, Z.; Rekik, I.; Yin, Z.; Ji, W.; Gao, Y.; Wang, Q.; Zhang, J.; Shen, D. Transformers in Medical Image Analysis. *Intell. Med.* **2023**, *3*, 59–78. <https://doi.org/10.1016/j.imed.2022.07.002>.
44. Bielecka, M.; Bielecki, A.; Obuchowicz, R.; Piórkowski, A. Universal Measure for Medical Image Quality Evaluation Based on Gradient Approach. In *Computational Science—ICCS 2020*; Krzhizhanovskaya, V.V., Závodszky, G., Lees, M.H., Dongarra, J.J., Sloat, P.M.A., Brissos, S., Teixeira, J., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 12140, pp. 406–417. ISBN 978-3-030-50422-9.
45. Obuchowicz, R.; Piórkowski, A.; Urbanik, A.; Strzelecki, M. Influence of Acquisition Time on MR Image Quality Estimated with Nonparametric Measures Based on Texture Features. *BioMed Res. Int.* **2019**, *2019*, 3706581. <https://doi.org/10.1155/2019/3706581>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.