ORIGINAL RESEARCH

Improving the Robustness of Deep-Learning Models in Predicting Hematoma Expansion from Admission Head CT

Anh T. Tran, Gaby Abou Karam, Dorin Zeevi, Adnan I. Qureshi, Ajay Malhotra, Shahram Majidi, Santosh B. Murthy, Soojin Park, Despina Kontos, Guido J. Falcone, Kevin N. Sheth, and Seyedmehdi Payabvash

ABSTRACT

BACKGROUND AND PURPOSE: Robustness against input data perturbations is essential for deploying deep-learning models in clinical practice. Adversarial attacks involve subtle, voxel-level manipulations of scans to increase deep-learning models' prediction errors. Testing deep-learning model performance on examples of adversarial images provides a measure of robustness, and including adversarial images in the training set can improve the model's robustness. In this study, we examined adversarial training and input modifications to improve the robustness of deep-learning models in predicting hematoma expansion (HE) from admission head CTs of patients with acute intracerebral hemorrhage (ICH).

MATERIALS AND METHODS: We used a multicenter cohort of n=890 patients for cross-validation/training, and a cohort of n=684 consecutive ICH patients from two stroke centers for independent validation. Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) adversarial attacks were applied for training and testing. We developed and tested four different models to predict ≥3mL, ≥6mL, ≥9mL, and ≥12mL HE in independent validation cohort applying Receiver Operating Characteristics (ROC) Area Under the Curve (AUC). We examined varying mixtures of adversarial and non-perturbed (clean) scans for training as well as including additional input from the hyperparameter-free Otsu multi-threshold segmentation for model.

RESULTS: When deep-learning models trained solely on clean scans were tested with PGD and FGSM adversarial images, the average HE prediction AUC dropped from 0.8 to 0.67 and 0.71, respectively. Overall, the best performing strategy to improve model robustness was training with 5-to-3 mix of clean and PGD adversarial scans and addition of Otsu multi-threshold segmentation to model input, increasing the average AUC to 0.77 against both PGD and FGSM adversarial attacks. Adversarial training with FGSM improved robustness against similar type attack but offered limited cross-attack robustness against PGD-type images.

CONCLUSIONS: Adversarial training and inclusion of threshold-based segmentation as an additional input can improve deep-learning model robustness in prediction of HE from admission head CTs in acute ICH.

ABBREVIATIONS: ATACH-2= Antihypertensive Treatment of Acute Cerebral Hemorrhage; AUC= Area Under the Curve; Dice=Dice coefficient; CNN= Convolutional Neural Network; FGSM= Fast Gradient Sign Method; ICH= Intracerebral hemorrhage; HD= Hausdorff distance; HE= Hematoma expansion; PGD= Projected Gradient Descent; ROC= Receiver Operating Characteristics; VS= Volume similarity.

Received month day, year; accepted after revision month day, year.
From the Department of Radiology (A.T.T., D.Z., D.K., S. Payabvash) and Neurology (S. Park), NewYork-Presbyterian/Columbia University Irving Medical Center, Columbia University, New York, NY; Department of Radiology and Biomedical Imaging (G.A., A.M.) and Neurology (G.J.F., K.N.S.), Yale School of Medicine, New Haven, CT; Zeenat Qureshi Stroke Institute and Department of Neurology (A.I.Q.), University of Missouri, Columbia, MO; Department of Neurosurgery (S.M.), Icahn School of Medicine at Mount Sinai, Mount Sinai Hospital, New York, NY; and Department of Neurology (S.B.M.), Weill Cornell Medical College, Cornell University, New York, NY.

Seyedmehdi Payabvash and this study were supported by Doris Duke Charitable Foundation (2020097), NIH (K23NS118056), and NVIDIA Applied Research Accelerator Program.

Please address correspondence to Seyedmehdi (Sam) Payabvash, MD, Center for Innovation in Imaging Biomarkers and Integrated Diagnostics (CIMBID), Department of Radiology, Columbia University, 530 West 166th Street, 5th Floor, New York, NY; email: sp4479@columbia.edu; @SamPayabvash

SUMMARY SECTION

PREVIOUS LITERATURE: The success of deep-learning models in medical imaging relies on their generalizability and stability of predictions. However, small, imperceptible voxel-level perturbations (adversarial attacks) can mislead models, reducing their accuracy and raising concerns about their clinical reliability. Prior studies suggest that adversarial training (i.e. incorporating perturbed images in training set) can improve deep-learning model robustness; however, optimal strategies for increasing the robustness of hematoma expansion (HE) prediction remain unclear. Ensuring robustness in HE prediction models from admission head CTs is critical for their clinical use in guiding targeted treatment of at-risk hemorrhagic stroke patients.

KEY FINDINGS: Training with a mixture of perturbed and clean (non-perturbed) non-contrast head CT scans can effectively improve the robustness of deep-learning models for hematoma segmentation and HE prediction (classification). Additionally, incorporating input from hyperparameter-free Otsu threshold-based segmentation of head CTs can further increase the robustness of these models.

KNOWLEDGE ADVANCEMENT: We reported the optimal adversarial training strategy and the benefits of adding threshold-based Otsu segmentation to improve the robustness of hematoma segmentation and HE prediction deep-learning models. These models can guide targeted therapies in hemorrhagic stroke, and our proposed methodology can be extended to development of other robust deep-learning models.

INTRODUCTION

In addition to prediction accuracy, several factors determine the trustworthiness of deep-learning models in healthcare, including robustness, generalizability, interpretability, fairness, and security.[1] Robustness refers to model's ability to maintain performance despite noise or perturbations in input data.[2] In medical image analysis, a robust deep-learning model can accurately classify or segment clinical scans even in the presence of noise, corrupted voxels, or blurring. Main strategies to improve the robustness of deep-learning models include adversarial training,[2] preprocessing, and postprocessing techniques such as feature squeezing,[3] and model designs that can detect perturbed image inputs (i.e. adversarial attacks).[4]

Adversarial images are generated by intentional manipulation of original scans to include subtle, voxel-level perturbations designed to maximize prediction errors of deep-learning models.[1] These voxel-level changes are often imperceptible to human eyes and different from introducing random noise into images since adversarial images are purposely crafted to challenge model's prediction, and are more effective in evaluating models robustness against input perturbations.[1] The vulnerability of deep-learning models to adversarial images (attacks) raises concerns about their robustness and trustworthiness in real-world clinical practice.[5]

Adversarial training refers to including adversarial images in the training set and is one of the most computationally efficient strategies for enhancing the robustness of deep-learning models.[2] However, training models exclusively with adversarial images may reduce their accuracy on unperturbed inputs.[5, 6] In this study, we systematically applied and optimized adversarial training to improve the robustness of deep-learning models for predicting hematoma expansion (HE) from admission non-contrast head CT scans of patients with acute intracerebral hemorrhage (ICH). To further improve model robustness, we also incorporated an automated thresholding step during image processing. HE affects nearly one-third of ICH patients within six hours of admission and is an independent predictor of neurological deterioration, disability, and mortality.[7] Predicting HE is clinically valuable as it can guide targeted anti-expansion and hemostatic therapies in ICH.[8, 9] Recent studies have shown that radiomic features of the hematoma and deep-learning models can predict HE from admission non-contrast head CTs with greater accuracy than visual markers determined by expert reviewers, such as the blend sign and swirl sign.[10, 11] Improving the robustness of HE prediction models can make them more suitable for the real-world clinical practice.

MATERIALS AND METHODS

Patients' ascertainment

We used the admission and follow-up head CT scans from the Antihypertensive Treatment of Acute Cerebral Hemorrhage (ATACH-2) multicenter randomized trial for training and cross-validation.[12] Briefly, ATACH-2 evaluated intensive blood pressure reduction in patients presenting with spontaneous ICH and at least one systolic blood pressure >180 mmHg but found no treatment benefit.[12] The trial included 110 sites in the United States, Germany, Japan, China, Taiwan, and South Korea.[12] For external validation, we tested the models in an independent cohort of consecutive patients presenting with spontaneous ICH to Yale Health Stroke Centers (Yale New Haven Hospital at York Street and Saint Raphael Campuses) from January 2015 to December 2023. Patients were included if they had baseline and follow-up non-contrast head CT scans within 6, and 36 hours after the onset, respectively. Subjects with metal/streak artifacts or surgical interventions affecting hematoma lesions on either baseline or follow-up scan, precluding accurate segmentation of hematoma, were excluded from the analysis. The Institutional Review Boards of participating centers in ATACH-2 clinical trial approved follow-up analysis of data. Our retrospective study of ICH patients at Yale received separate Institutional Review Board approval.

Generation of ground truth labels for HE

We manually segmented hematoma lesions on baseline and follow-up head CTs, as described previously.[10, 11, 13] Segmentations were performed by trained research associates and then reviewed/revised by a board-certified neuroradiologist with over 10 years of experience. The intra- and inter-rater reliability of segmentations were determined in a subset of scans using intra-class correlation, ranging from 0.92 to 0.94.[10, 11, 13] We trained and validated separate models for prediction of ≥3mL, ≥6mL, ≥9mL, and ≥12mL HE from baseline to follow-up scan as binary classifications.[14] These HE thresholds were described originally to predict poor outcomes with increasing specificity and positive predictive values in ICH patients.[14]

Backbone of the HE prediction model

Our image analysis pipeline included preprocessing steps for skull removal,[11] adjustment to brain window-level, and co-registration of brain to isotropic 3D template,[15] as detailed in supplemental material. Our prior experience showed that a dual input from CT slices and hematoma mask will improve the classification performance for HE prediction.[11] For the backbone of hematoma segmentation model, we used nnUNET,[16] a versatile U-Net-shaped convolutional neural network (CNN) with self-configuration capabilities.[16] For classification, we modified the DenseNet121 3D CNN for prediction of HE. We developed and tested four different models for prediction of ≥3mL, ≥6mL, ≥9mL, and ≥12mL HE. The overall model structure is depicted in Figure 1 and further described below.

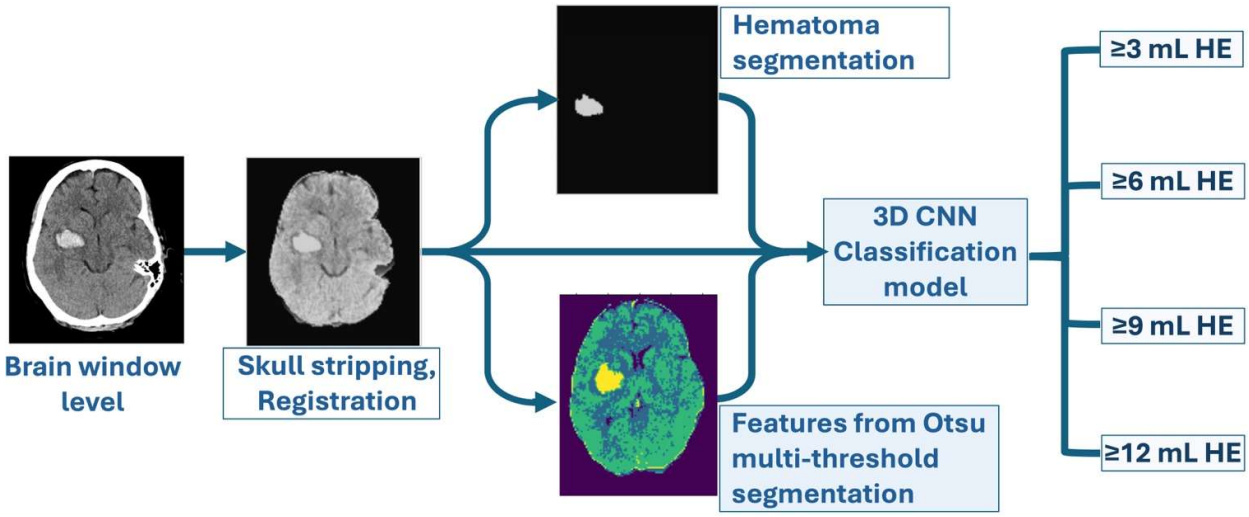

FIG 1. Our final model included preprocessing of head CTs with adjusting the images to brain window level (80:40 Hounsfield Unites), skull removal, and registration to a similar size 128×128×128 template. The model included (3, 128, 128, 128) inputs from skull-stripped head CT slices, automated hematoma segmentation masks, and Otsu multi-threshold segmentation. We trained and tested four separate model for prediction of ≥3mL, ≥6mL, ≥9mL, and ≥12mL HE.

Creating adversarial Images

In this study, we applied Fast Gradient Sign Method (FGSM)[17] and Projected Gradient Descent (PGD)[18] to create adversarial images from brain CT. These methods are designed to maximize classification error of the deep-learning model's while minimizing the difference between the adversarial and original images. All adversarial image generation methods are bounded under a predefined perturbation size ε, which represents the maximum change added to each pixel/voxel value, after normalizing their intensities to 0-to-1 range. Details of the two methods are included in supplemental material, and examples are depicted in Figure 2.

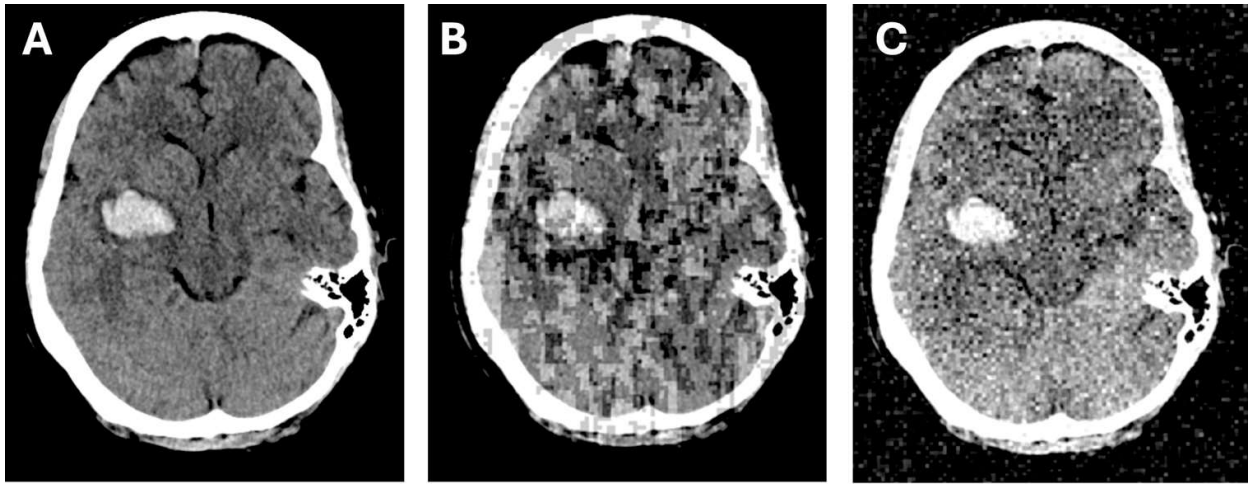

FIG 2. Examples of (A) an original head CT slice, and (B) adversarial images after applying Fast Gradient Sign Method (FGSM),[17] and (C) Projected Gradient Descent (PGD),[18] with ε=0.1 perturbation size.

Adversarial Training to Improve Model Robustness

Inclusion of adversarial images in training dataset can improve the resilience of deep-learning models to adversarial attacks, by encouraging the models to learn more robust and meaningful features, rather than overfitting to cohort-specific patterns of the training dataset. However, training of models exclusively with adversarial images can reduce their accuracy for unperturbed input. By exposing the model to both adversarial and original (clean) images during training, the deep-learning model will learn to classify both image types more accurately compared to being trained only on one image type. In this study, we trained the model with varying mixtures of clean-to-adversarial images (from 5:5 to 5:1) to identify the optimal strategy for maximum accuracy. Clean images included original scans, and basic augmentations such as resize, rotate, and flip without adversarial perturbations. Adversarial images included perturbed variations of original scans and basic augmentations.

Otsu multi-threshold segmentation to improve robustness

The Otsu thresholding method is a popular technique in image segmentation due to its simplicity and efficiency.[19] The algorithm automatically separates voxels of input images into predefined number of classes based on their signal intensity by maximizing the between-class variance.[19] Since such multi-class segmentations maintain the overall spatial consistency of the image, their outputs are more resilient to subtle, inconsistent changes introduced by adversarial attacks. Multi-class segmentation of input images can improve the robustness of final deep-learning models and counter adversarial attacks.[20] The task of distinguishing between multiple classes at a pixel/voxel level forces the model to become more discriminative. This heightened class-wise differentiation reduces the impact of adversarial attacks, which tend to confuse the deep-learning models by blending features of different classes. In addition, multi-threshold segmentation often requires capturing overlapping features between different classes. This redundancy can act as a form of defense because perturbations that affect one part of the image may not be sufficient to fool the entire segmentation map. The model can use redundancy to cross-verify predictions, making it harder for an adversarial attack to succeed. We added Otsu multi-threshold segmentation to both segmentation and classification pipelines. Thus, the inputs for the classification CNN were concatenation of brain images, hematoma segmentation mask, and the features from Otsu thresholding multi-class segmentation. In our exploratory analysis, we found that four-level thresholding has better delineation of hemorrhage in brain tissue based on average Dice results (Figure 3).

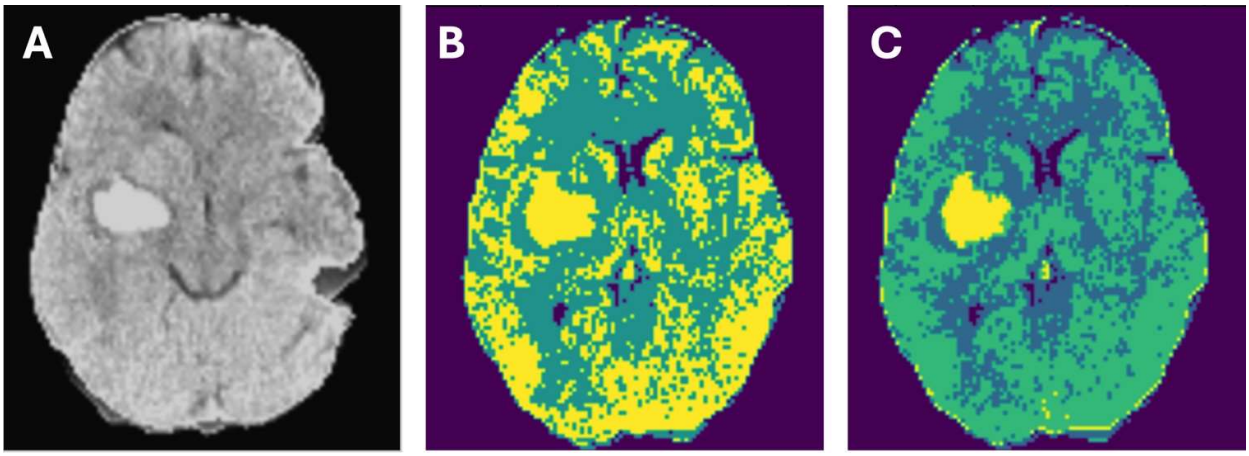

FIG 3. Example of (A) the skull-stropped and registered brain CT, with (B) three-level and (C) four-level Otsu thresholding multi-class segmentation. The results of four-level thresholding had better overlap with hematoma lesions.

Training and validation of hematoma segmentation model:

We used five-fold cross-validation for training the hematoma segmentation model. The base model was the nnUNet version 2.0, and we included Otsu multi-threshold segmentation as an additional input to head CT slices. For training we applied weight_decay= 3e-5, initial learning rate=0.001, num_epochs=100, PolyLRScheduler, optimizer=SGD, and patch window size= (96, 96, 128). These hyperparameters were optimized through a grid search strategy. The model accuracy was determined using Dice, Hausdorff Distance (HD), and Volume Similarity (VS) as described in supplemental material. Online Figure 1 represents an example of loss function and Dice diagram during training process. Robustness of segmentation model was evaluated by mean Dice, HD, and VS in independent test cohorts without and with adversarial images of various perturbation sizes.

Training and validation of HE prediction models:

For training of HE prediction models, we used similar five-fold cross-validation split as segmentation model, applying DenseNet121 from MONAI platform with loss_function= BCEWithLogitsLoss(), optimizer= Adam(), scheduler= ReduceLROnPlateau, learning rate= 0.001, weight_decay=1e-4, augmentation=(RandFlip, RandZoom, RandRotate), batch_size=10. These hyperparameters were optimized through a grid search strategy. The inputs (3, 128, 128, 128) of our final model were skull-stripped brain CT slices, automated hematoma segmentation masks, and Otsu four-class thresholding segmentation. We developed and tested separate models for prediction of ≥3mL, ≥6mL, ≥9mL, and ≥12mL HE. We tested model prediction performance using the Area Under the Curve (AUC) of Receiver Operating Characteristics (ROC) analysis. Online Figure 2 represents an example of loss function and AUC changes during training process. We compared the effectiveness of the proposed strategies in improving model robustness by testing their prediction AUC in clean (non-perturbed) and perturbated images.

RESULTS

Patient characteristics:

Of 1000 patients enrolled in multicentric ATACH-2 trial, n=890 were included in cross-validation/training cohort, 11 were excluded due to CT and surgical hardware artifacts, 59 due to missing follow-up head CT scans, and 40 due to missing clinical information. From 940 ICH patients in Yale stroke registry, n=684 were included in independent validation cohort, 10 excluded due to CT and surgical hardware artifacts, 97 due to missing baseline or follow-up head CT scans, and 149 due to missing clinical information. Details of patients' characteristics and CT scan information are listed in Table 1. Overall, patients in independent validation cohort were more likely to be female, have older age, larger baseline and follow-up hematoma volumes (all $p<0.001$).

Table 1: The demographic and clinical characteristics of patients in training/cross-validation versus independent test cohort.

Column A	Training/cross-validation from ATACH-2 (n=890)	Independent Test (n=684)	p-value
Sex – Male	543 (60.8%)	358 (54.8%)	<0.001
Age* [years]	62.15 ± 13.1	69.7 ± 14.3	<0.001
Hypertension	706 (79.3%)	583 (85.2%)	0.17
Diabetes	172 (19.3%)	181 (26.5%)	0.28
Hyperlipidemia	221 (24.8%)	332 (48.5%)	0.81
Atrial fibrillation	30 (3.4%)	150 (21.9%)	0.17
NIHSS score at baseline			0.86
0-4	148 (16.57%)	243 (35.57%)	
5-9	235 (26.31%)	119 (17.40%)	
10-14	241 (26.98%)	81 (11.85%)	
15-19	161 (18.02%)	92 (13.46%)	
20-25	71 (7.95%)	52 (7.60%)	
>25	37 (4.17%)	30 (4.39%)	
Unknown		66 (9.73%	
Baseline hematoma volume [mL]	12.9 ± 12.6	18.7 ± 20.6	<0.001
Follow-up hematoma volume [mL]	15.6 ± 16.6	22.9 ± 25.8	<0.001
CT voxel Spacing [mm]	[0.46 ± 0.09, 0.47 ± 0.09]	[0.46 ± 0.04, 0.46 ± 0.04]	
Slice thickness [mm]	5.20 ± 1.86	4.81 ± 0.69	
Min axial [n x n]	[418 x 418]	[472 x 472]	
Max axial [n x n]	[512 x 734]	[1024 x 1024]	
Number of slices	31.2 ± 18.0	35.0 ± 11.2	

Data is presented as number (percentage) or mean ± standard deviation. Nominal variables are compared using chi square and continuous variables are compared using t test.

Hematoma segmentation model

Online Table 1 provides details of hematoma segmentation performance across different model structures and with adversarial training. Using clean (non-perturbed) CT scans for training with the baseline nnUNET model, we achieved an average Dice score of 0.91±0.14 for hematoma segmentation in clean CT scans. However, the Dice of model trained on clean CT images dropped to 0.24±0.29, and 0.71±0.19 when tested on FGSM and PGD adversarial images (ε=0.1). By adding Otsu multi-level thresholding to model input, the average Dice for segmentation on FGSM and PGD adversarial images improved to 0.51±0.37 and 0.81±0.17, respectively, while maintaining a similar average Dice score in clean CTs. Further improvements in adversarial performance were achieved by training with a one-to-one mixture of clean and adversarial images using segmentation model with additional input from Otsu. Training with a mixture of clean scans and FGSM adversarial images yielded average Dice scores of 0.83±0.20 and 0.73±0.19 when tested on FGSM and PGD adversarial images, respectively. Training with a mixture of clean scans and PGD adversarial images resulted in average Dice scores of 0.83±0.20 and 0.85±0.24 when tested on FGSM and PGD adversarial images, respectively. Segmentation performance on clean CT scans was maintained, with an average Dice score of 0.91.

Improving robustness of HE prediction models

Online Table 2 summarizes the performance of different HE prediction models with and without adversarial training. Figure 4 presents the performance of various mixtures of clean and adversarial images in training when tested on FGSM and PGD adversarial images with the highest perturbation level applied in our study (ε=0.1). The average AUCs of HE prediction models using inputs from brain CT slices and hematoma segmentations were 0.8 on clean scans, dropping to 0.67 and 0.71 when tested on FGSM and PGD adversarial images (ε=0.1). Notably, the hematoma masks were generated via automated segmentation by a model optimized with adversarial training. Including Otsu multi-threshold segmentation in the prediction model improved the average AUC on FGSM adversarial images (ε=0.1) to 0.74 (Online Table 2). Overall, the PGD adversarial attacks with ε values of 0.01 and 0.1 reduced model performance more than FGSM attacks. Adversarial training to each specific attack type improved model robustness against similar attack – i.e. a mixture of FGSM and clean images in training improved models' AUC when tested against FGSM attacks more effectively than against PGD, and vice versa. However, there was a trend indicating that PGD adversarial training also provided some improvement against FGSM attacks, whereas FGSM adversarial training offered limited performance gains against PGD attacks (Figure 4). Overall, a 5:3 mixture of clean and PGD adversarial images yielded the highest average AUC in HE prediction against both PGD and FGSM adversarial attacks. Any combination of clean and adversarial images during training preserved the baseline average AUC of 0.8 in HE prediction when tested on clean scans.

DISCUSSION

Successful application of deep-learning models in clinical practice depends on their generalizability and stability against perturbations in input data. Several recent studies have reported the vulnerability of medical imaging classification and segmentation models to adversarial attacks or small perturbations in input data.[5, 21, 22] We found that subtle voxel-level perturbations negatively impact the performance of both ICH segmentation and HE prediction models. Then, we showed that adversarial training with a mix of clean and perturbed scans can improve the robustness of hematoma segmentation and HE prediction models while maintaining accuracy in clean datasets. We also found cross-attack robustness of hematoma segmentation and HE prediction model against FGSM attacks after training on PGD adversarial images. From a computational efficiency perspective, our findings suggest that training on a mix of clean and a subset of adversarial images (e.g. a 5-to-3 ratio) can sufficiently enhance model robustness against various types of input perturbations. In addition, we showed the benefit of including Otsu multi-threshold segmentation as an additional input to improve both segmentation and classification

robustness. The Otsu algorithm provides a simple, efficient, and parameter-free method by automatically selecting the optimal threshold to maximize between-class variance in the segmented image.[19, 20] Overall, the key takeaway from our findings is twofold: first, to include a subset of adversarial images as part of data augmentation during model training; and second, to incorporate hyperparameter-free, threshold-based segmentation as an additional input for the model. These strategies may not increase model accuracy on clean (non-perturbed) test cohorts, but they can improve the model's robustness and stability in handling noisy and distorted images.

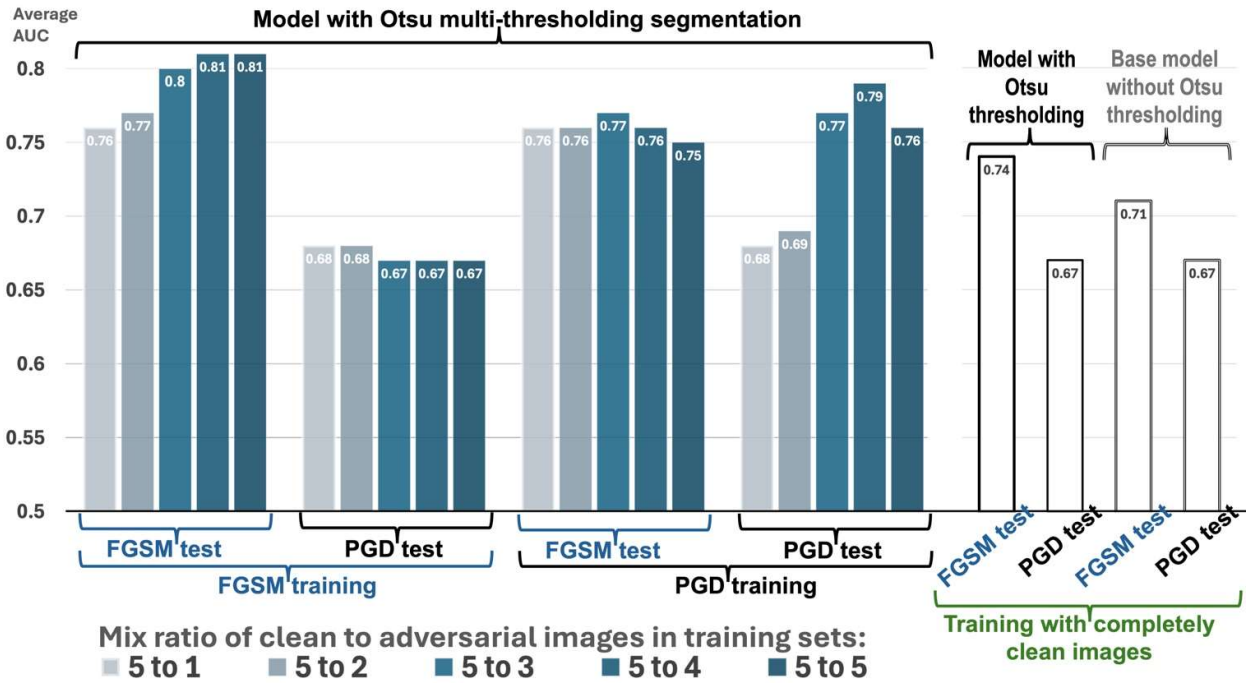

FIG 4. Different mix ratios of FGSM and PGD adversarial images with clean (non-perturbed) scans were used to train HE prediction models (with model input including brain CT slices, automatically segmented hematoma masks, and Otsu multi-level thresholding). The average ROC AUC of ≥3mL, ≥6mL, ≥9mL, and ≥12mL HE prediction models are shown, when tested against adversarial attacks with ε=0.1 perturbation size. PGD attacks were more effective in reducing the AUCs than FGSM attack. On the other hand, PGD training also improved model performance against FGSM attacks. Notably, addition of Otsu-thresholding in prediction model slightly improved model performance against FGSM attacks even in the absence of any adversarial training.

Identifying ICH patients at risk of HE is crucial for guiding hemorrhagic stroke treatment. As an independent and potentially modifiable risk factor for poor outcomes and mortality, HE has been the target of anti-expansion and hemostatic therapies.[12, 23-25] Recent clinical trials showed the benefits of intensive blood pressure reduction as part of a critical care bundle or in pre-hospital setting.[8, 9] Post-hoc analyses of prior clinical trials also showed improved outcomes and reduced HE after ultra-early blood pressure reduction, despite negative overall trial results.[26-28] This highlights the need for reliable identification of ICH patients at risk for HE, which can expand the eligibility of patients to receive intensive therapies, and expedite potential inter-hospital transfer decisions in acute settings. Recent studies have shown that radiomics and deep-learning models can provide more reliable, automated prediction of HE.[10, 11, 29-31] However, for clinical practice use, these models require extensive external validation to confirm generalizability and robustness. Our findings provide a framework for improving the robustness of HE prediction models, paving the way for their potential clinical application.

Clean data might not fully represent the diverse range of inputs a deep-learning model could encounter in real-world scenarios. Robustness examines the potential failure of models when exposed to perturbed data points. Adversarial images represent potential inputs that are sparsely represented in the clean dataset, making accurate predictions challenging for the model. Notably, ambiguous and outlier images cannot be modeled by simply adding noise, as adversarial attacks are specifically crafted to induce prediction failures.[1] This makes them better suited for evaluating model behavior under extreme inputs. Adversarial training acts as a form of regularization, helping to prevent overfitting and improve model stability. By exposing the model to both clean and adversarial examples – as part of data augmentation during training – the model learns to balance the trade-off between accuracy on clean data and robustness to adversarial perturbations, resulting in more stable predictions.[1, 2, 21] As also shown in our study, features learned during adversarial training are often transferable across different types of adversarial attacks. This suggests that even if the attacks encountered during testing differ from those encountered in training, the model may still leverage the knowledge gained from training to defend against other forms of adversarial attacks.

Prior studies suggest that adversarial training can improve deep-learning model accuracy, especially in underrepresented examples. Liu et al. reported that the addition of adversarial images and adversarial synthetic nodules to the training data improved model robustness in detecting underrepresented lung nodules and resistance to noise perturbations in low-dose CT scans.[32] Vatian et al. showed that the addition of adversarial images can improve classification accuracy of chest CTs with lung nodules and brain MRIs with gliomas.[33] Hu et al. found that adversarial training improves the generalizability and accuracy of deep learning models in the classification of prostate cancer MRIs.[34] However, an analysis of a large X-ray dataset (n=22,433) suggests no improvement in deep-learning model performance with adversarial training when trained with a large sample size.[35] In our study, there was no significant change in model performance on the clean test

cohort with training on a mix of clean and adversarial images. However, the primary goal of adversarial training is to maintain model performance when exposed to examples not included in the existing dataset, as evidenced by the robustness of adversarially trained models against attacks in both segmentation and classification tasks.

We found that subtle voxel-level perturbations (e.g. ε=0.001 or 0.01) had a limited impact on the accuracy of hematoma segmentation and HE prediction models, whereas moderate perturbations (ε = 0.1) significantly diminished the performance of both segmentation and prediction models. This is consistent with prior studies showing a gradual decrease in the performance of medical image analysis models when ε>0.01.[5, 21, 22] Notably, perturbations with ε>0.1 typically become visually conspicuous, rendering such images beyond the scope of model robustness improvement.[5, 21, 22] In addition, the impact of FGSM attacks on segmentation model performance was greater than that of PGD. This is likely because PGD, being an iterative method, refines its perturbations to maximize misclassification over multiple steps, resulting in more localized perturbations than the broader signal changes introduced by FGSM. Thus, when attacking segmentation models, PGD's local adjustments can retain some structural integrity, leading to better Dice scores compared to FGSM attack.

The Otsu method is a hyperparameter-free image segmentation technique that automatically identifies the optimal threshold(s) to maximize between-class variance of voxels in segmentation classes.[19, 20] For the Otsu algorithm, the number of threshold levels must be predefined, and our experiments showed that a four-level threshold on skull-stripped brain CT slices achieves the best overlap with hematoma segmentation masks. We found that adding Otsu multi-threshold segmentation results as feature extraction in the model input improved the robustness of both hematoma segmentation and HE prediction models. This is likely because clustering different intensity levels within the image enhances the features of underlying pathologies (like hematoma) and reduces the influence of noise. With multi-level thresholding, the impact of minor perturbations is diminished, as the image is segmented into broader intensity-based regions. Thus, Otsu's method can act as a denoising layer, reducing the model's sensitivity to random noise and adversarial changes.[36]

There is inherent differences in computational efficiency of adversarial versus standard training process. Standard training is computationally more efficient, without generating additional adversarial images as part of data augmentation. This results in faster training times and reduced hardware requirements, making standard training ideal for resource-constrained settings. By contrast, adversarial training requires generating perturbed examples through techniques such as FGSM or PGD. These methods increase the computational load due to additional forward and backward passes, leading to significantly longer training times and the need for high-performance servers. On the other hand, inference efficiency typically remains similar for both training approaches, but the benefits of adversarial training may extend to improved model confidence and performance in challenging scenarios. For settings with limited computational resources, hybrid strategies, such as introducing adversarial examples gradually through cumulative learning, can help balance robustness and training efficiency. These approaches reduce the overall computational burden while still enhancing model robustness. In our analysis, adversarial training increased the training time by 3 times compared to standard method; whereas, testing of each patient took 35±10 seconds depending on the number of slices for either model type.

The strengths of our study include the use of a multicenter dataset for training and an independent validation cohort. We also explored different mix ratios in adversarial training and demonstrated the benefits of multi-threshold segmentation in enhancing the robustness of deep-learning models. Aside from adversarial training, other methods to improve deep-learning robustness include conventional data augmentation to expand training datasets, ensemble learning boosting or voting, regularization to prevent overfitting, feature squeezing in pre- or postprocessing steps,[3] and identification of perturbed image inputs.[4] In addition, approaches such as transfer learning, domain adaptation, cross-validation, and ensemble learning can improve generalization and model adaptation across different datasets. In our study, aside from adversarial training, strategies such as data augmentation, regularization, dropouts, and normalization contributed to model robustness and generalizability. Future studies can combine these methods to further improved deep-learning model robustness and generalizability, with the ultimate goal of maintaining model performance across different centers in clinical practice.

Our study also has several limitations. Our training cohort was limited by the inclusion and exclusion criteria of the ATACH-2 trial. The benefits of Otsu multi-level thresholding may also be limited to ICH, given the inherent contrast between hematoma and unaffected brain tissue, and may not generalize to other clinical scenarios. Other defense mechanisms against adversarial attacks, such as feature squeezing or adversarial sample detection, were not evaluated in this study. Although focusing on the brain window/level setting was the optimal choice for HE prediction based on our prior experience, other window/level settings may enable the extraction of additional CNN features. Exploring hybrid approaches or multi-window preprocessing in future iterations could further enhance the model's optimization for clinical applications. It should be noted that the primary goal of improving deep-learning model robustness is to maintain performance stability against image perturbations caused by data heterogeneity and noise; whereas, addressing major metal or motion artifacts requires dedicated artifact reduction models.

CONCLUSIONS

We found that hematoma segmentation and HE prediction models trained exclusively on non-perturbed (clean) head CT scans are vulnerable to voxel-intensity modifications from adversarial attacks. Our findings show that training with a mix of adversarial and clean datasets, along with incorporating threshold-based segmentation of brain CTs as additional input, can improve model robustness against adversarial attacks. Additionally, training on one type of adversarial images can provide cross-attack robustness against other types. Overall, our results suggest that including adversarial images as a subset of data augmentation instances in the training process, along with incorporating hyperparameter-free thresholding as an additional input, can improve the robustness of classification and segmentation deep-learning models. Increased robustness of deep-learning models will lead to higher stability when exposed to variations in input data, such as noise, artifacts, and visually imperceptible perturbations.

REFERENCES

1. Paschali M, Conjeti S, Navarro F, et al. Generalizability vs. Robustness: Adversarial Examples for Medical Imaging. 2018:arXiv:1804.00504
2. Apostolidis KD, Papakostas GA. A Survey on Adversarial Deep Learning Robustness in Medical Image Analysis. Electronics 2021;10:2132

3. Xu W, Evans D, Qi Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. 2017:arXiv:1704.01155
4. Li X, Zhu D. Robust Detection of Adversarial Attacks on Medical Images. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 2020:1154-1158
5. Joel MZ, Umrao S, Chang E, et al. Using Adversarial Images to Assess the Robustness of Deep Learning Models Trained on Diagnostic Images in Oncology. JCO Clin Cancer Inform 2022;6:e2100170
6. Rodriguez D, Nayak T, Chen Y, et al. On the role of deep learning model complexity in adversarial robustness for medical images. BMC Med Inform Decis Mak 2022;22:160
7. Abou Karam G, Chen MC, Zeevi D, et al. Time-Dependent Changes in Hematoma Expansion Rate after Supratentorial Intracerebral Hemorrhage and Its Relationship with Neurological Deterioration and Functional Outcome. Diagnostics (Basel) 2024;14
8. Li G, Lin Y, Yang J, et al. Intensive Ambulance-Delivered Blood-Pressure Reduction in Hyperacute Stroke. N Engl J Med 2024;390:1862-1872
9. Ma L, Hu X, Song L, et al. The third Intensive Care Bundle with Blood Pressure Reduction in Acute Cerebral Haemorrhage Trial (INTERACT3): an international, stepped wedge cluster randomised controlled trial. Lancet 2023;402:27-40
10. Haider SP, Qureshi AI, Jain A, et al. Radiomic markers of intracerebral hemorrhage expansion on non-contrast CT: independent validation and comparison with visual markers. Front Neurosci 2023;17:1225342
11. Tran AT, Zeevi T, Haider SP, et al. Uncertainty-aware deep-learning model for prediction of supratentorial hematoma expansion from admission non-contrast head computed tomography scan. NPJ Digit Med 2024;7:26
12. Qureshi AI, Palesch YY, Barsan WG, et al. Intensive Blood-Pressure Lowering in Patients with Acute Cerebral Hemorrhage. N Engl J Med 2016;375:1033-1043
13. Haider SP, Qureshi AI, Jain A, et al. The coronal plane maximum diameter of deep intracerebral hemorrhage predicts functional outcome more accurately than hematoma volume. Int J Stroke 2022;17:777-784
14. Dowlatshahi D, Demchuk AM, Flaherty ML, et al. Defining hematoma expansion in intracerebral hemorrhage: relationship with patient outcomes. Neurology 2011;76:1238-1244
15. Rorden C, Bonilha L, Fridriksson J, et al. Age-specific CT and MRI templates for spatial normalization. Neuroimage 2012;61:957-965
16. Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203-211
17. Ian J. Goodfellow JS, Christian Szegedy. Explaining and Harnessing Adversarial Examples. arXiv 2014:1412.6572
18. Aleksander Madry AM, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv 2017:1706.06083v06084
19. Otsu N. A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics 1979;9:62-66
20. Jing Z, Tang B. Improved image segmentation method based on Otsu thresholding and level set techniques. Journal of Physics: Conference Series 2024;2813:012017
21. Bortsova G, Gonzalez-Gonzalo C, Wetstein SC, et al. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. Med Image Anal 2021;73:102141
22. Muoka GW, Yi D, Ukwuoma CC, et al. A Comprehensive Review and Analysis of Deep Learning-Based Medical Image Adversarial Attack and Defense. Mathematics 2023;11:4272
23. Gladstone DJ, Aviv RI, Demchuk AM, et al. Effect of Recombinant Activated Coagulation Factor VII on Hemorrhage Expansion Among Patients With Spot Sign-Positive Acute Intracerebral Hemorrhage: The SPOTLIGHT and STOP-IT Randomized Clinical Trials. JAMA Neurol 2019;76:1493-1501
24. Meretoja A, Yassi N, Wu TY, et al. Tranexamic acid in patients with intracerebral haemorrhage (STOP-AUST): a multicentre, randomised, placebo-controlled, phase 2 trial. Lancet Neurol 2020;19:980-987
25. Naidech AM, Grotta J, Elm J, et al. Recombinant factor VIIa for hemorrhagic stroke treatment at earliest possible time (FASTEST): Protocol for a phase III, double-blind, randomized, placebo-controlled trial. Int J Stroke 2022;17:806-809
26. Leasure AC, Qureshi AI, Murthy SB, et al. Intensive Blood Pressure Reduction and Perihematomal Edema Expansion in Deep Intracerebral Hemorrhage. Stroke 2019;50:2016-2022
27. Leasure AC, Qureshi AI, Murthy SB, et al. Association of Intensive Blood Pressure Reduction With Risk of Hematoma Expansion in Patients With Deep Intracerebral Hemorrhage. JAMA Neurol 2019;76:949-955
28. Li Q, Warren AD, Qureshi AI, et al. Ultra-Early Blood Pressure Reduction Attenuates Hematoma Growth and Improves Outcome in Intracerebral Hemorrhage. Ann Neurol 2020;88:388-395
29. Teng L, Ren Q, Zhang P, et al. Artificial Intelligence Can Effectively Predict Early Hematoma Expansion of Intracerebral Hemorrhage Analyzing Noncontrast Computed Tomography Image. Front Aging Neurosci 2021;13:632138
30. Chen Q, Zhu D, Liu J, et al. Clinical-radiomics Nomogram for Risk Estimation of Early Hematoma Expansion after Acute Intracerebral Hemorrhage. Acad Radiol 2021;28:307-317
31. Ma C, Zhang Y, Niyazi T, et al. Radiomics for predicting hematoma expansion in patients with hypertensive intraparenchymal hematomas. Eur J Radiol 2019;115:10-15
32. Liu S, Setio AAA, Ghesu FC, et al. No Surprises: Training Robust Lung Nodule Detection for Low-Dose CT Scans by Augmenting With Adversarial Attacks. IEEE Trans Med Imaging 2021;40:335-345
33. Vatian A, Gusarova N, Dobrenko N, et al. Impact of Adversarial Examples on the Efficiency of Interpretation and Use of Information from High-Tech Medical Images. 2019 24th Conference of Open Innovations Association (FRUCT); 2019:472-478
34. Hu L, Zhou DW, Guo XY, et al. Adversarial training for prostate cancer classification using magnetic resonance imaging. Quant Imaging Med Surg 2022;12:3276-3287
35. Han T, Nebelung S, Pedersoli F, et al. Advancing diagnostic performance and clinical usability of neural networks via adversarial training and dual batch normalization. Nat Commun 2021;12:4315
36. Anastasiou T, Karagiorgou S, Petrou P, et al. Towards Robustifying Image Classifiers against the Perils of Adversarial Attacks on Artificial Intelligence Systems. Sensors (Basel) 2022;22

SUPPLEMNENTARY MATERIALS

Preprocessing of head CTs

- **Skull stripping**: First, we removed voxels with intensities <0 and >200 Hounsfield Units (HU) to facilitate skull stripping. After removing much of the skull and subcutaneous fat tissue using these HU thresholds, we applied mathematical dilation and erosion operations based on brain morphology to exclude any remaining soft tissue from the scalp and face.
- **Adjusting the brain window and level**: When reviewing medical scans, radiologists usually use window width and level to accentuate the contrast between target tissues. For example, the optimal setting for visual inspection of brain tissue on non-contrast CT scans is a window level=40, and width=80, which we applied in our pre-processing step. In our prior experiments (reference 11 of the article) we found that brain window series can optimally predict hematoma expansion from admission head CT compared to using no window, with no improvement from addition of other head CT window levels proposed of subdural hematoma or stroke evaluation.
- **Registration**: We perform registration of complete 3D images to an isotopic 1.5 mm CT template with (128, 128, 128) dimensions using the ANTsPy library.

Evaluation of segmentation model performance

- **Dice Similarity Coefficient**: Dice measures the volumetric overlap between segmentation results and ground truth. Dice is computed where A is the set of foreground voxels in the ground truth and B is the corresponding set of foreground voxels in the segmentation result.

$$\text{Dice} = \frac{2(\text{A} \cap \text{B})}{|\text{A}| + |\text{B}|} \quad (1)$$

- **Hausdorff distance (HD):** HD is a measure of the maximum distance of a (segmentation) mask to the nearest point in the other (ground truth) mask [1], and is defined as

$$\text{d}_\text{H}(\text{X},\text{Y}) = \max\{\text{d}_{\text{XY}}, \text{d}_{\text{YX}}\} = \max\left\{\max_{\text{x}\in\text{X}} \min_{\text{y}\in\text{Y}} \text{d}(\text{x},\text{y}), \max_{\text{y}\in\text{Y}} \min_{\text{x}\in\text{X}} \text{d}(\text{x},\text{y}),\right\} \quad (2)$$

- **Volume Similarity (VS)**: VS measures and compares the absolute volume of the segmented result and ground truth, is defined as

$$VS = 1 - \frac{|v1 - v2|}{v1 + v2} \quad (3)$$

Adversarial methods

- **Fast Gradient Sign Method (FGSM)**: FGSM is the earliest gradient-based model proposed for generating adversarial images (reference 17 of the article). The single-step FGSM attack perturbs the original image by a fixed amount along the direction (sign) of the gradient of adversarial loss. Given input image x, perturbation size ε, loss function J, and target label $(y_1, \dots, y_n)$, the adversarial image can be computed as:

$$x_{adv} = x + \varepsilon sign(\nabla_x J(x, (y_1, \dots, y_n)))$$

- **Projected Gradient Descent (PGD)**: PGD method creates stronger gradient attacks than FGSM by initializing adversarial examples at random points within the allowed range and then running multiple iterations (reference 18). PGD iteratively perturbs the input with a smaller step size. After each iteration,

the updated adversarial example is projected onto the ε-ball centered around the original input and then clipped to remain within a permitted range:

$$x^t = \Pi_\varepsilon(x^{t-1} + \alpha sign(\nabla_x J(x^t, (y_1, \dots, y_n))))$$

Code sharing

https://github.com/payabvashlab/robustnessHE

Online Table 1. The results of original model and our proposed model evaluating adversarial attacks on hematoma segmentation.

Model	Images in training	Adversarial attacks in testing	Permutation ε	Dice	HD	VS
Base model (nnUNET)	Clean	Clean		0.91±0.14	3.38±8.76	0.94±0.12
		FGSM	0.001	0.91±0.18	3.98±15.74	0.94±0.15
			0.01	0.86±0.20	6.81±24.68	0.90±0.18
			0.1	0.24±0.29	39.14±40.81	0.37±0.31
		PGD	0.001	0.91±0.14	3.38±8.88	0.94±0.12
			0.01	0.90±0.18	4.48±16.58	0.94±0.15
			0.1	0.71±0.19	8.50±12.79	0.74±0.19
Our model (nnUNet + Otsu multi-threshold segmentation)	Clean	Clean		0.91±0.14	3.55±9.43	0.94±0.12
		FGSM	0.001	0.91±0.14	3.75±11.13	0.94±0.12
			0.01	0.90±0.16	5.01±19.39	0.93±0.18
			0.1	0.51±0.37	32.04±49.24	0.65±0.32
		PGD	0.001	0.91±0.14	3.65±10.13	0.94±0.12
			0.01	0.90±0.16	4.85±15.85	0.94±0.14
			0.1	0.81±0.17	7.11±17.70	0.88±0.15
Our model (nnUNet + Otsu multi-threshold	Clean + FGSM (1:1)	Clean		0.91±0.14	3.34±10.01	0.94±0.12
		FGSM	0.001	0.91±0.14	3.69±10.11	0.94±0.12

segmentation)			0.01	0.90±0.14	3.75±10.20	0.94±0.12
			0.1	0.83±0.20	5.52±9.89	0.90±0.15
		PGD	0.001	0.91±0.14	3.72±10.13	0.94±0.11
			0.01	0.90±0.14	3.78±10.85	0.94±0.11
			0.1	0.73±0.19	6.76±11.8	0.76±0.19
	Clean + PGD (1:1)	Clean		0.91±0.14	3.46±9.11	0.94±0.11
		FGSM	0.001	0.90±0.15	3.99±10.91	0.93±0.14
			0.01	0.90±0.15	4.25±11.20	0.93±0.14
			0.1	0.83±0.20	5.52±11.89	0.90±0.15
		PGD	0.001	0.91±0.14	3.78±10.12	0.94±0.13
			0.01	0.90±0.15	4.04±10.53	0.93±0.13
			0.1	0.85±0.24	14.72±15.31	0.74±0.19

We used nnUNET as the backbone of the hematoma segmentation model. Our proposed model included nnUNET and Otsu multi-thresholding segmentation. Training cohorts included the original head CTs or 1-to-1 mixtures of clean and adversarial images. We applied the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) with ε=0.001, 0.01, 0.1 perturbation size. The models were tested on clean, FGSM, and PGD adversarial images. The mean ± standard deviation of Dice coefficient, Hausdorff Distance (HD), and Volume Similarity (VS) of segmentation results in the independent validation cohort are reported.

Online Figure 1: An example of loss and Dice diagram during training/cross-validation of segmentation model

Online Figure 2. An example of loss diagram and Receiver Operating Characteristics (ROC) Area Under Curve (AUC) during training/cross-validation of hematoma expansion (HE) prediction model