

Adversarial attack driven data augmentation for medical images

Mst. Tasnim Pervin¹, Linmi Tao², Aminul Huq³

¹Department of Computer Science, American International University, Dhaka, Bangladesh

²Department of Computer Science and Technology, Tsinghua University, Beijing, China

³Department of Computer Science and Engineering, Brac University, Dhaka, Bangladesh

Article Info

Article history:

Received Oct 6, 2022

Revised Apr 14, 2023

Accepted Apr 24, 2023

Keywords:

Adversarial machine learning

Adversarial examples

Medical image segmentation

Data augmentation

UNet

ABSTRACT

An important stage in medical image analysis is segmentation, which aids in focusing on the required area of an image and speeds up findings. Fortunately, deep learning models have taken over with their high-performing capabilities, making this process simpler. The deep learning model's reliance on vast data, however, makes it difficult to utilize for medical image analysis due to the scarcity of data samples. Too far, a number of data augmentations techniques have been employed to address the issue of data unavailability. Here, we present a novel method of augmentation that enabled the UNet model to segment the input dataset with about 90% accuracy in just 30 epochs. We describe the usage of fast gradient sign method (FGSM) as an augmentation tool for adversarial machine learning attack methods. Besides, we have developed the method of Inverse FGSM, which improves performance by operating in the opposite way from FGSM adversarial attacks. In comparison to the conventional FGSM methodology, our strategy boosted performance up to 6% to 7% on average. The model became more resilient to hostile attacks because to these two strategies. An innovative implementation of adversarial machine learning and resilience augmentation is revealed by the overall analysis of this study.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Mst. Tasnim Pervin

Department of Computer Science, American International University

Dhaka, Bangladesh

Email: tasnim.pervin@aiub.edu

1. INTRODUCTION

With the aid of progressive and efficient computational power, in-depth network architecture, and a variety of discriminating tasks related to Computer Vision applications, such as regression, prediction, image segmentation, or object recognition, the substantial structures of deep learning models have been demonstrated to be very effective [1]–[3]. Deep learning networks, however, seldom function as intended in the absence of a sizable dataset. This restriction turns out to be considerably more significant for the field of medical image processing since access to a vast amount of data is not like tossing a coin. Many supervised biomedical segmentation techniques concentrate on hand-engineered preparation procedures and structures to address these issues [4], [5]. To expand the amount of training instances, hand-tuned data augmentation is also frequently used [6], [7]. Additionally, the human-level labeling of the dataset's images by medical professionals is exceedingly costly and labor-intensive, and it is also the cause of the substantial variances in resolution and noise in tissue appearance [8]. The gap between the validation and training parts can be minimized by using data augmentation as a solution to this issue. Overfitting can be overcome with the

addition of new data. In some situations, augmentation functions such random nonlinear deformations or picture rotations are simple to use and successful at increasing segmentation accuracy [4], [5], [7], [9]. These functions, however, can be quite sensitive to the parameter selection and have a limited capacity to mimic genuine fluctuations [10]. The categorization of skin lesions by Esteva *et al.* [11] and Litjens *et al.* [12] the classification of liver lesions, among other papers, have reports of this phenomenon. The work of data augmentation and the use of adversarial machine learning were both completed in this study. Recent research demonstrates that adversarial attacks are also capable of impairing the performance of segmentation models. However, these conventional augmentation strategies are unable to make these models resistant to various attack methods. Inception-v3 and DenseNet-121, two high-performing deep learning models, were shown by Bortsova *et al.* [13] to lose strength when subjected to fast gradient sign method (FGSM) and projected gradient descent (PGD) assaults for three distinct types of datasets (ophthalmology, radiology, and pathology). Paschali *et al.* [14] additionally examined how adversarial attacks affected the effectiveness of the segmentation (SegNet, UNet, DenseNet) and classification (Inception V3, Inception V4, MobileNet) models. Even while adversarial attacks are intended to reduce model performance, we employed an attack method called FGSM in this case to increase the dataset for our benefit. We have also presented a fresh viewpoint on how to improve FGSM further by changing its operational methodology. The following is a list of this paper's main contributions:

- In contrast to the typical goal of weakening models, in our study we employed adversarial machine learning in support of deep learning models. To construct an adversarial sample for the aim of augmentation while limiting overfitting, we employed the FGSM attack approach.
- We bring up an innovative strategy called Inverse-FGSM, which aims to minimize loss by adjusting input data. Positive sounds were added to the model in place of adversarial noises, which enhanced performance.
- Unless they have been trained to withstand attacks from adversaries or employ some sort of defense mechanism, all deep learning models are often weak to attacks. In order to make the model more resilient to future comparable attacks, we applied adversarial training-based augmentation of adversarial images to the original set in our research.

Later sections are separated into smaller units like background research on segmentation models, data augmentation, and attack models of adversarial machine learning is compiled in section 2. Section 3 provides a thorough discussion of the techniques that were employed. The dataset is described in section 4 along with an analysis of the effects of data augmentation. Section 5 concludes the essay towards the end.

In order to prevent overfitting, medical image segmentation models are modified to have fewer convolutional blocks than convolutional models due to the scarcity of training samples. The most well-known segmentation network, UNet, was developed by Ronnerberger *et al.* [7] it produces exact localized higher resolution segmented pictures and can be trained on a small number of training images. Novikov *et al.* [15] suggested a tweak to UNet where they employed larger feature maps but with fewer parameters altering AL-dropout architectures to prevent overfitting. Hwang and Park [16] suggested a model called network-wise training of convolutional networks (NWCN) that uses a multi-stage training technique for distilled segmentation outputs with smooth boundary for the consistent usage of contextual information together with appropriate resolution. Another study by Sarker *et al.* [17] focused on sharp-edged, segmented dermoscopic pictures, and to reach this goal, they utilized log likelihood and end point error loss. By including some look-alike data, data augmentation tries to increase the size of the original training sample. In the study on skin lesion classification, Esteva *et al.* [11] showed that deep convolutional networks are extremely effective when used with a bigger dataset for medical picture analysis. This has advanced the usage of convolutional neural networks (CNNs) for related tasks including liver lesion categorization, brain scan image processing, and many more [12]. It takes a lot of work to capture pictures, especially for technologies like magnetic resonance imaging (MRI) and computed tomography (CT). The lack of patient cooperation, the rarity of the illnesses, and the absence of human competence make it extra harder. Many studies on data augmentation have been done in an effort to reduce these problems. Modern data augmentation methods may be categorized into two groups: i) geometric transformation that flips, changes the color space, crops, rotates, or introduces noise and ii) deep learning-based (generative adversarial networks (GANs), neural style transfer, and feature space translation). Elastic deformations, patch extraction, and adjusting red, green, blue (RGB) channel intensities are only a few of the techniques used for picture categorization [18], [19]. These methods essentially involve changing the level of a picture, such as by scaling, rotation, cropping, and in-depth modifications. The more recent method of data augmentation encompasses a variety of intricately modified algorithms that have been developed in a wide range of fields, including emotion classification, text recognition from scene, text localization, and human position detection [20]–[22].

Modern machine learning (ML) and deep learning (DL) models were able to deliver accurate results that were above any reasonable expectations because to improved processing capabilities and high

configuration architecture. Recent developments in adversarial machine learning, however, have this illusion. Simply challenging a competitive high-performing model with adversarial cases might cause it to misbehave. A sample of input that causes a model to make the incorrect prediction is referred to as an adversarial example. Adversarial examples, which are deliberately prepared using unnoticeable noise injections, are used in these attack tactics to corrupt ML/DL models. The security of practical applications is seriously threatened by adversarial instances. Even an adversarial scenario that seeks to deceive model M1 frequently succeeds in deceiving model M2 all along [23]. By using the transferability criterion of adversarial examples, a system may be attacked even if all parameter values are unknown. Engstrom *et al.* [24] deep CNN models may be made to perform misclassification by assaulting them with a few simple geometric modifications. Each dataset's performance was decreased as a result of random changes (a reduction of 26% for MNIST, 72% for CIFAR1, and 28% for ImageNet). In a different study, Ian Goodfellow and his coauthors proposed the FGSM, which generates 89.4 percent misclassification with 97.6% confidence and uses a maxout network to create adversarial samples [25]. Su *et al.* [26] developed the one-pixel attack, which may result in 70.97% of input photos being misclassified by modifying just one pixel. Every conventional ML/DL model is stated as $F(x) = y$ if the provided dataset is $(x; y)$, where x represents input, y represents label, and F is considered to be a classifier which translates input data to relevant labels. However, for adversarial machine learning, the classifier F also receives input together with a small amount of perturbation δ in limited by a threshold, $0 < \delta \leq \epsilon$, rendering it impossible to predict original labels. The transformation is expressed as $F(x + \delta) \neq y$. A resilient model ought to be able to effectively deal with this imperceptible disruption.

2. PROPOSED METHOD

The fundamental concept is to enrich the tiny dataset for training with a new set of training examples produced using adversarial machine learning techniques in order to develop segmentation or classification algorithms. In the conventional approach, adversarial algorithms concentrate on the non-robust attributes of the pictures to reduce the performance of the model, but we use this strategy in the other direction to strengthen weak areas of images with fewer details. This method provides the models with extra input knowledge. Figure 1 graphic seeks to depict the entire process: i) use the original dataset to train the model for segmenting localized images; ii) attack the model using an adversarial attack method that uses model architecture and input gradients to create perturbations; and iii) add adversarial examples to the original samples to create a new training set before feeding the model for training. This can improve the outcomes of segmentation and make the simulation more resistant to attacks. This stands in stark contrast to the conventional augmentation methods formerly outlined. Although adversarial augmentations might not be instances that are likely to appear in the test set, they can strengthen areas where the learnt decision boundary is vulnerable. A novel idea that has not been well investigated and evaluated is the usefulness of adversarial training in the manner of augmentation. Although it has been demonstrated that using adversarial examples to inject noise improves performance, its use in achieving the goal of decreasing over-fitting has not been established.

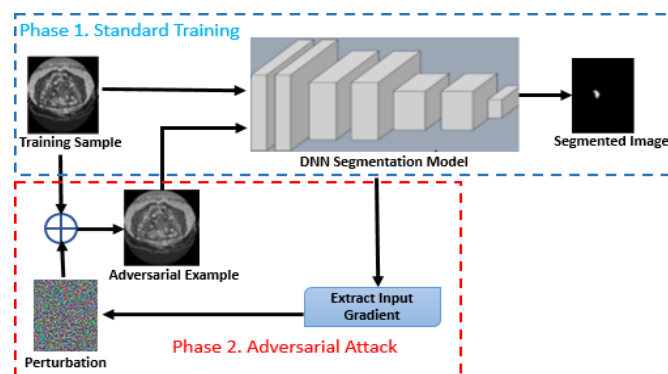


Figure 1. Workflow of proposed approach

3. METHOD

3.1. UNet

For the semantic segmentation of medical pictures, the UNet model by Ronnerberger *et al.* [7] is highly well-liked. Labels and high-resolution pictures with each of the pixels designating a particular class

are also produced as a result of segmentation. The models are composed of an expanding/decoder path and a contraction/encoder path. The difference between the encoder and decoder paths is that the encoder path just consists of a mass of convolutional and maxpooling layers, while the decoder path calls for certain transposed convolutional layers to pursue the position of the context, or the “where” information, of the picture. As the depth increases in the encoder route, the dimension of the pictures steadily decreases, going from $128 \times 128 \times 3$ to $8 \times 8 \times 256$. The size of the picture steadily rises with decreasing depth in the decoder route, for example, from $8 \times 8 \times 256$ to $128 \times 128 \times 1$, which aids in producing input similar segmented resultant visuals. At each level, the feature maps from the encoder and features from transposed convolutional layers of decoder are combined for a particular localization. The architecture is shaped like a U as a result of this feature. UNet’s advantage is that it only has fully convolutional layers and no thick layers, allowing it to take input pictures of any size.

3.2. Fast gradient sign method (FGSM)

We want to use adversarial machine learning approaches for improved segmentation, as per the approach. Here, we choose Goodfellow *et al.* [25] fast gradient sign method (FGSM), the quickest adversarial technique. The steps involved that has been followed are: i) utilize input image (x) consisting of ground mask/label (y), ii) generate estimations using a machine learning algorithm and determine the loss ($L(\theta, x, y)$), iii) estimate the gradient (∇_x) of the loss w.r.t. input and identify the sign of that gradient, iv) develop noise pattern employing perturbation (σ) along with the sign of the gradient, and v) create adversarial examples x_{adv} combining noise pattern/perturbations with input that maximizes loss. These actions result in the creation of adversarial samples with similar input that can trick neural networks into making the wrong predictions. We may infer how FGSM enhances the volume of loss throughout adversarial example production from Figure 2(a). Due to the model’s goal of locating the highest loss following the gradient, this occurs. The mathematical form of this process is shown in (1). This approach is intended to be used for augmentation. This raises the question of how these hostile instances, which are intended to deceive the model, can function as an augmentation dataset. To do this, we want to employ adversarial training as an adversarial augmentation. Adversarial training will increase the model’s resilience and make it aware of assaults. According to (2), it is based on resolving a min-max optimization issue. The outer minimization locates model parameters with minimal loss on adversarial cases, whereas the inner maximization locates the largest classification loss. As a result, by adjusting the perturbation, σ (where $0 < \sigma \leq \epsilon$), we get examples that are positive. The goal is to improve performance by supplementing these instances using adversarial training rather than utilizing maximum perturbation (ϵ).

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y)) \quad (1)$$

$$\overbrace{\min_{\theta} \frac{1}{N} \sum_{i=1}^N \min_{x_i - x_i^0 \leq \sigma} L(h_{\theta}(x_i), y_i)}^{\text{outer-minimization}} \quad (2)$$

inner-maximization

3.3. Inverse FGSM

The term of FGSM served as the inspiration for this strategy. Instead of changing weights to lessen the loss, FGSM aims to change the inputs to a maximum loss through backpropagation of gradients. The inverse FGSM (InvFGSM) variation of the FGSM technique is what we suggest. In order to improve performance, our suggested InvFGSM technique utilizes an inverse strategy that involves modifying inputs to a minimized loss. Only the gradient’s sign counts for adversarial noise when using the FGSM method. The inverse of the sign gives a positive adversarial noise which will help in loss minimization, much as the sign of the derivatives produces adversarial noises that are mixed with input to produce adversarial instances. The performance of the model may be improved by adding these adversarial cases with positive sounds to the data. The mathematical form of this approach is seen in (3). We may infer from Figure 2(b) that the InvFGSM approach helps to keep loss minimization rather than growing as in FGSM. With this, we may produce advantageous adversarial cases for a fresh set of augmentations. To enhance these instances, we take a similar technique as outlined in the preceding section. However, since we use a reverse strategy to locate gradients with the least amount of loss during the development of positive adversarial examples, this training may be thought of as a min-min optimization issue, according to (4). The outer minimization seeks to identify model parameters while also reducing loss on those adversarial cases. The inner minimization seeks to find the smallest classifying loss during the creation of adversarial examples throughout the domain of greatest perturbation, $0 < \sigma \leq \epsilon$. This training strategy is intended to be used for adversarial augmentation.

$$x_{adv} = x - \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y)) \quad (3)$$

$$\overbrace{\min_{\theta} \frac{1}{N} \sum_{i=1}^N \underbrace{\min_{x_i - x_i^0 \leq \sigma} L(h_{\theta}(x_i), y_i)}_{\text{inner-minimization}}}_{\text{outer-minimization}} \tag{4}$$

4. RESULTS AND DISCUSSION

4.1. Dataset acquisition

Two different datasets (colon cancer and lumbar-CT) were utilized for this project. The Computer Vision and Graphics lab, in partnership with Beijing Tsinghua Changgung Hospital and Peking Union Medical College Hospital, has made both datasets available. The dataset for colon cancer includes 1285 CT scan pictures and a comparable amount of ground mask images that show the required localization of cancer cells. Nearly 370 CT scan pictures and an even proportion of ground mask images of the lumbar vertebral body are included in the Lumbar-CT collection. There are training, validation, and testing sets for each dataset. To speed up processing, all of the photos were downsized from their initial resolution of 512×512×3 to 128×128×3 using the Jittor platform [27].

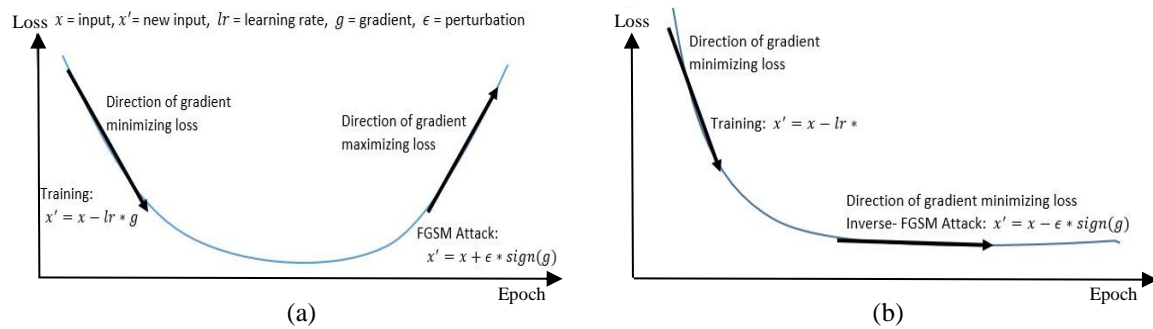


Figure 2. Effects of (a) FGSM and (b) InvFGSM on the loss curve in relation to model parameter

4.2. Experimental design and model development

We repeated experimenting multiple times with parameter tuning to find out most suitable setting. The following are the parameters for both model training with and without data augmentation: categorical cross-entropy loss function, SGD as optimizer, momentum is 0.99, learning rate is 0.1, and batch size is 16. Batch size and learning rate can be increased for faster convergence if proper GPU Power is available.

4.3. Augmentation of data through adversarial training

The impact of the suggested augmentation approach will be covered in this section. Table 1 shows the difference in the segmentation accuracy that was achieved before and after augmentation. The UNet model typically segments medical images pretty effectively. Up to epoch 30, UNet provides roughly 80% accurate segmentation. Without a doubt, data augmentation can make these results superior. Adversarial machine learning is a relatively recent method of augmentation that we used. Compared to UNet alone, FGSM and UNet perform around 3% better. However, as expected, inverse FGSM boosts this performance even more, demonstrating an increase of roughly 6%.

Table 1. The models' performance throughout different epochs

Models	Mean IoU over epochs							
	0 th	Colon cancer dataset			Lumbar-CT dataset			
		10 th	20 th	30 th	0 th	10 th	20 th	30 th
U-Net	0.4955	0.5586	0.7311	0.8084	0.5004	0.9535	0.9781	0.9829
U-Net + FGSM	0.4987	0.7839	0.8129	0.8394	0.6161	0.8186	0.8938	0.9387
U-Net + InvFGSM	0.6159	0.8264	0.8643	0.8986	0.6737	0.9027	0.9486	0.9774

4.4. Effect of increased perturbation on model robustness

Adversarial machine learning attack strategies can be added to strengthen the model. In Table 2, this phenomena is illustrated. To test the model's resilience, we used random perturbation with a range of 0 to 0.2. It is obvious how the model changes when the noise level rises. Even 0.2 epsilon noise reduces segmentation performance by up to 35%. Contrarily, adversarial training unquestionably assisted the model in regaining

performance and robustness. As optical imperceptibility is another goal in addition to strong performance for adversarial training, the degree of perturbation was set at 0.1 because a larger rate of perturbation results in visibly distorted adversarial pictures. For the Colon Cancer and Lumbar CT datasets, respectively, Figures 3 and 4 show some examples of Figure 3(a) and 4(a) original input, Figure 3(b) and 4(b) ground truth that has been used to generate, Figure 3(c) and 4(c) adversarial image for augmentation using proposed InvFGSM method which helps better segmentation producing, Figure 3(d) and 4(d) predicted mask for Colon Cancer and Lumbar CT dataset respectively.

Table 2. Effect of attacks on the model robustness with varying perturbation (ϵ) for dataset (a) colon cancer and (b) lumbar CT

Attacking model	Adversarial training	Mean IoU		
		$\epsilon=0.0012$	$\epsilon=0.0118$	$\epsilon=0.1176$
FGSM(a)	Before	0.7860	0.7246	0.6028
	After	0.9071	0.8242	0.7720
Inv-FGSM(a)	Before	0.7749	0.5841	0.5467
	After	0.8724	0.8729	0.8477
FGSM(b)	Before	0.8630	0.7007	0.4944
	After	0.9106	0.8812	0.9150
Inv-FGSM(b)	Before	0.9262	0.8517	0.4953
	After	0.9806	0.9624	0.9169

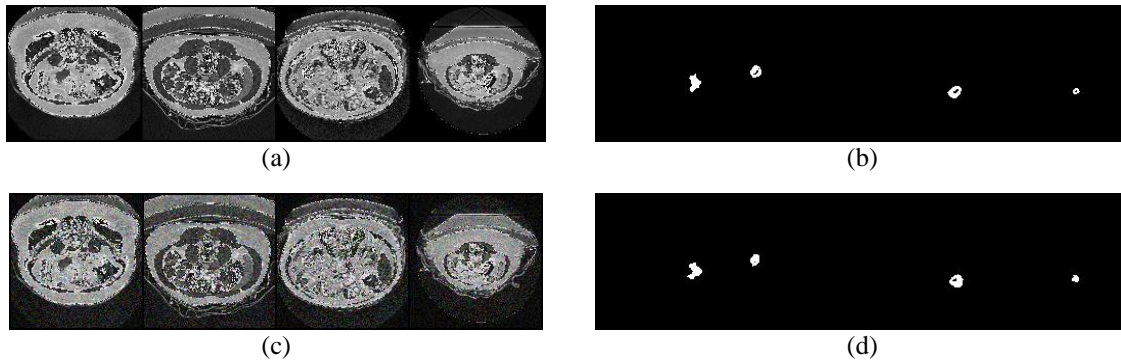


Figure 3. Segmentation performance on the colon cancer dataset using adversarial augmentation by InvFGSM: (a) original input, (b) ground truth/mask, (c) adversarial image, and (d) predicted mask after augmentation

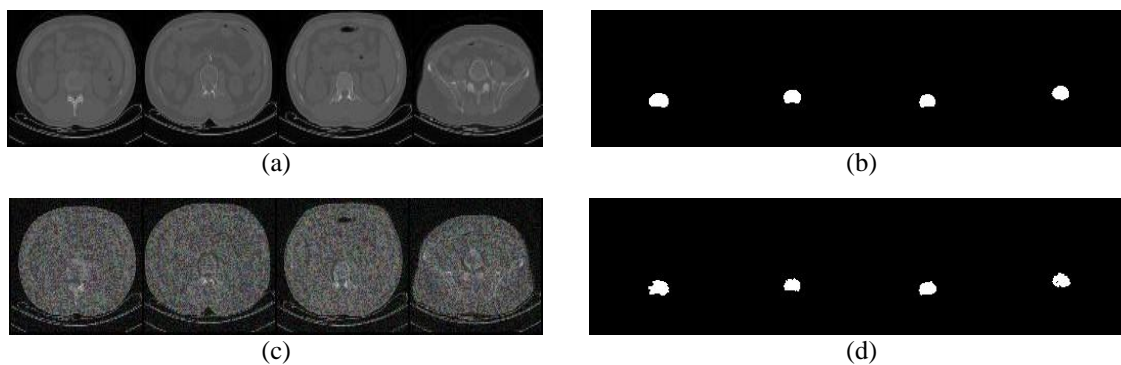


Figure 4. Segmentation performance on the Lumbar-CT dataset using adversarial augmentation by InvFGSM: (a) original input, (b) ground truth/mask, (c) adversarial image, and (d) predicted mask after augmentation

4.5. Comparison to traditional augmentations

We also compared our method's performance against several commonly used data augmentation techniques. The results can be found in Table 3. We compared our method against flipping, rotating, cropping, color jittering and GAN. Amongst all the methods our proposed InvFGSM performed the best. The second-best result was achieved by using GAN method for data augmentation.

Table 3. Performance of proposed method against different methods

Model	Colon Cancer	Lumbar CT
U-Net	0.8084	0.9829
U-Net + Flipping	0.8127	0.8055
U-Net + Rotating	0.8135	0.8131
U-Net + Cropping	0.8215	0.8753
U-Net + Color Jitter-	0.7989	0.8069
U-Net + GAN	0.8455	0.821
U-Net + FGSM	0.8394	0.9387
U-Net + InvFGSM	0.8986	0.9774

4.6. Discussion

This study aims to combine the subject of medical image analysis with utilization of adversarial machine learning. We used one adversarial attack technique to see if it might be used to supplement medical data. Because medical pictures are often low-resolution, even little changes can have an impact on the ability to identify diseases. We have to use the attack strategy with additional caution because of this. In order to aid in augmentation without detracting from the model, we used a relatively straightforward attack method called FGSM and tweaked the parameter. Later, we put our own approach for augmentation to use. This method was influenced by the theoretical underpinnings of FGSM. However, there are still a number of effective assault methods that may have more pronounced consequences for augmentation or, alternatively, the result might be quite the reverse. Future research into this problem will help us make a more accurate assumption about how to use adversarial attacks for augmentation.

5. CONCLUSION

Using a dataset of colon cancer cases and CT scan images of the lumbar vertebrae, we attempted to deploy a deep learning algorithm enabling cancer cell segmentation in this study. We also developed a novel method for data augmentation for improved segmentation. Data augmentation is seen as a vital step since limited data results in the overfitting issue when utilized with deep learning models. Although it cannot serve as a complete backup for incomplete data in cases when there is no class sample, it can be used to identify overfitting. As a novel method of enhancing the data, we explored with adversarial machine learning attack strategies, which were successful in enhancing the efficiency of segmentation for this dataset. Even while model robustness is the traditional goal of adversarial machine learning, this study clearly demonstrates that it can also be utilized for data augmentation tasks. Although further research is needed to determine whether this benefits all models equally, it should be seen as a promising beginning for a novel method of data enrichment employing adversarial machine learning.




REFERENCES

- [1] P. N. Druzhkov and V. D. Kustikova, "A survey of deep learning methods and software tools for image classification and object detection," *Pattern Recognition and Image Analysis*, vol. 26, no. 1, pp. 9–15, Jan. 2016, doi: 10.1134/S1054661816010065.
- [2] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021, doi: 10.1109/TPAMI.2021.3059968.
- [3] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0197-0.
- [4] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. N. L. Benders, and I. Išgum, "Automatic segmentation of MRI brain images with a convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1252–1261, May 2016, doi: 10.1109/TMI.2016.2548501.
- [5] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1240–1251, May 2016, doi: 10.1109/TMI.2016.2538465.
- [6] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, 2003, vol. 1, pp. 958–963, doi: 10.1109/ICDAR.2003.1227801.
- [7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Prepr. arXiv1505.04597*, May 2015.
- [8] K. K. Leung *et al.*, "Robust atrophy rate measurement in Alzheimer's disease using multi-site serial MRI: Tissue-specific intensity normalization and parameter selection," *NeuroImage*, vol. 50, no. 2, pp. 516–523, 2010, doi: 10.1016/j.neuroimage.2009.12.059.
- [9] H. R. Roth *et al.*, "DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *MICCAI 2015: Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 556–564, doi: 10.1007/978-3-319-24553-9_68.
- [10] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 8535–8545, doi: 10.1109/CVPR.2019.00874.
- [11] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.
- [12] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.




- [13] G. Bortsova *et al.*, “Adversarial attack vulnerability of medical image analysis systems: Unexplored factors,” *Medical Image Analysis*, vol. 73, Oct. 2021, doi: 10.1016/j.media.2021.102141.
- [14] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, “Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples,” in *MICCAI 2018: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 2018, pp. 493–501, doi: 10.1007/978-3-030-00928-1_56.
- [15] A. A. Novikov, D. Lenis, D. Major, J. Hladuvka, M. Wimmer, and K. Buhler, “Fully convolutional architectures for multiclass segmentation in chest radiographs,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 8, pp. 1865–1876, Aug. 2018, doi: 10.1109/TMI.2018.2806086.
- [16] S. Hwang and S. Park, “Accurate lung segmentation via network-wise training of convolutional networks,” in *DLMIA 2017, ML-CDS 2017: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017, pp. 92–99, doi: 10.1007/978-3-319-67558-9_11.
- [17] M. M. K. Sarker *et al.*, “SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks,” in *MICCAI 2018: Medical Image Computing and Computer Assisted Intervention*, 2018, pp. 21–29, doi: 10.1007/978-3-030-00934-2_3.
- [18] D. C. Cireřan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, “High-performance neural networks for visual object classification,” *Prepr. arXiv.1102.0183*, Feb. 2011.
- [19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
- [20] A. Gupta, A. Vedaldi, and A. Zisserman, “Synthetic data for text localisation in natural images,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2315–2324, doi: 10.1109/CVPR.2016.254.
- [21] J. Yu, D. Farin, C. Kruger, and B. Schiele, “Improving person detection using synthetic training data,” in *2010 IEEE International Conference on Image Processing*, Sep. 2010, pp. 3477–3480, doi: 10.1109/ICIP.2010.5650143.
- [22] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, “Emotion classification with data augmentation using generative adversarial networks,” in *PAKDD 2018: Advances in Knowledge Discovery and Data Mining*, 2018, pp. 349–360, doi: 10.1007/978-3-319-93040-4_28.
- [23] C. Szegedy *et al.*, “Intriguing properties of neural networks,” *Prepr. arXiv.1312.6199*, Dec. 2013.
- [24] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, “A rotation and a translation suffice: Fooling CNNs with simple transformations,” in *ICLR 2019 Conference*, 2018, pp. 1–21.
- [25] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *Prepr. arXiv.1412.6572*, 2014.
- [26] J. Su, D. V. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, Oct. 2019, doi: 10.1109/TEVC.2019.2890858.
- [27] S.-M. Hu, D. Liang, G.-Y. Yang, G.-W. Yang, and W.-Y. Zhou, “Jitter: a novel deep learning framework with meta-operators and unified graph execution,” *Science China Information Sciences*, vol. 63, no. 12, Dec. 2020, doi: 10.1007/s11432-020-3097-4.

BIOGRAPHIES OF AUTHORS






Mst. Tasnim Pervin    is a PhD student at Department of Computer Science and Engineering in University of Nevada, Reno, United States of America. She has completed her Master’s in Computer Science and Technology from Tsinghua University in 2021. She completed her Bachelor’s from Rajshahi University of Engineering and Technology in 2017. She has served as a Lecturer at American International University-Bangladesh. Her research interest lies in the field of medical image analysis, adversarial machine learning and domain adaptation. He can be contacted at email: tasnim.pervin@aiub.edu.



Linmi Tao    holds a PhD in optical communications from University of Oxford, United Kingdom. She is an Associate Professor at Tsinghua University, China. He has completed his Ph.D. in Computer Science, Tsinghua University, Beijing, China, 2001 and Masters in Cognitive Science, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China, 1991. He is the Chair of ACM SIGCHI China from 2007 and was Program committee Co-Chair of VINCI in 2010. His research interest lies in computer vision, digital image/video processing and human computer interaction. She can be contacted at email: linmi@tsinghua.edu.cn.



Aminul Huq    is a PhD student at Department of Computer Science and Engineering in University of Nevada, Reno, United States of America. He received his Master’s in Computer Science and Technology from Tsinghua University in 2021. He completed his Bachelor’s from Rajshahi University of Engineering and Technology in 2017. He has served as a Lecturer at Brac University. His research interest lies in the field of multi-task learning and adversarial machine learning. He can be contacted at email: aminul.huq@bracu.ac.bd.