Adversarial-Robust Transfer Learning for Medical Imaging via Domain Assimilation

Xiaohui Chen and Tie Luo(✉)

Computer Science Department, Missouri University of Science and Technology, Rolla, MO 65409, USA
{xcqmk,tluo}@mst.edu

Abstract. Extensive research in Medical Imaging aims to uncover critical diagnostic features in patients, with AI-driven medical diagnosis relying on sophisticated machine learning and deep learning models to analyze, detect, and identify diseases from medical images. Despite the remarkable accuracy of these models under normal conditions, they grapple with trustworthiness issues, where their output could be manipulated by adversaries who introduce strategic perturbations to the input images. Furthermore, the scarcity of publicly available medical images, constituting a bottleneck for reliable training, has led contemporary algorithms to depend on pretrained models grounded on a large set of natural images—a practice referred to as transfer learning. However, a significant *domain discrepancy* exists between natural and medical images, which causes AI models resulting from transfer learning to exhibit heightened *vulnerability* to adversarial attacks. This paper proposes a *domain assimilation* approach that introduces texture and color adaptation into transfer learning, followed by a texture preservation component to suppress undesired distortion. We systematically analyze the performance of transfer learning in the face of various adversarial attacks under different data modalities, with the overarching goal of fortifying the model's robustness and security in medical imaging tasks. The results demonstrate high effectiveness in reducing attack efficacy, contributing toward more trustworthy transfer learning in biomedical applications.

Keywords: Medical images · natural images · transfer learning · colorization · texture adaptation · adversarial attacks · robustness · trustworthy AI

1 Introduction

Since its inception, Artificial Intelligence (AI) has evolved into a powerful tool across various domains. Particularly in the realm of medical diagnosis and treatment, AI has demonstrated impressive performance in predicting a range of diseases such as cancer, often treated as a classification problem. Beyond diagnosis,

D.-N. Yang et al. (Eds.): PAKDD 2024, LNAI 14648, pp. 335–349, 2024.
https://doi.org/10.1007/978-981-97-2238-9_26

applying AI to medical problems for object detection and segmentation has also become focal points for researchers. The year 2018 marked a milestone toward real-world integration of AI in clinical diagnosis, where IDX-DR became the first FDA-approved AI algorithm designed for automatic screening of Diabetic Retinopathy. On the research arena, numerous algorithms have been proposed that showcase remarkable performance in disease detection and diagnosis across diverse patient data modalities, such as MRI, CT, and Ultrasound. However, the lack of publicly available medical data, often attributed to privacy and the high cost of human expert annotation, remains a critical challenge for advancing medical AI research. In the meantime, model performance heavily depends on the size, quality, and diversity of training data to extract meaningful patterns. To address this challenge, researchers have turned to *transfer learning*, which takes models pretrained on large, extensive datasets of natural images and then fine-tunes the models' weights on the (smaller) medical datasets.

However, natural and medical images have inherent differences from each other, forming a gap that is largely overlooked when applying transfer learning. This gap, which we refer to as "domain discrepancy", encompasses two particular aspects that eventually lead to heightened *vulnerability* of medical AI models to adversarial attacks. First, the *monotonic biological textures* in medical images tend to mislead deep neural networks to paying extra attention to (larger) areas *irrelevant* to diagnosis. Second, medical images typically have simpler features than natural images, yet applying overparameterized deep networks to learn such simple patterns can result in a sharp loss landscape, as empirically observed in [19]. Both large attention regions and sharp losses render medical models trained via transfer learning susceptible to *adversarial examples*, the most common attack by adding small, imperceptible perturbations to original input images to induce significant changes in model output, resulting in mispredictions.

In this study, we introduce a novel approach called *Domain Assimilation* to bridge the gap between medical images and natural images. The aim is to align the characteristics of medical images more closely with those of natural images, thereby enhancing the robustness of models against adversarial attacks without compromising the accuracy of unaltered transfer learning. Our contributions are as follows:

- We propose a novel domain assimilation approach embodied by a texture-color-adaption module integrated into transfer learning. This module transforms medical images to resemble natural images more closely, thereby reducing domain discrepancy and enhancing model robustness against attacks.
- To prevent over-adaptation, which can lead to the loss of essential information in medical images and result in misdiagnoses, we introduce a novel Gray-Level Co-occurrence Matrix (GLCM) loss into the training process. This new loss incorporates texture preservation into the optimization process, which is instrumental to ensuring integrity of medical data and reliable diagnoses.
- We conduct extensive experiments across multiple modalities, including MRI, CT, X-ray, and Ultrasound, and evaluate the performance under various

adversarial attacks, such as FGSM, BIM, PGD, and MIFGSM. Our results demonstrate that the proposed texture-color adaption with GLCM loss effectively enhances the robustness of transfer learning while maintaining competitive model accuracy.

2 Related Work

The inherent differences between medical and natural images pose challenges and some efforts were invested to narrow their disparities [16,22]. A typical technique is grayscale image colorization, which is a prominent area in computer vision aiming to transform single-hue images into vibrant, colorful representations to enhance details and information for various applications. This field has garnered significant interest in the research community, finding applications in historical image restoration, architectural visualization, and the conversion of black-and-white movies into colorful ones, among others. Numerous existing works address colorization, either leveraging human-provided input such as scribbles or text, or utilizing color information from reference images [7,13]. Existing methods aim to predict the values of channels within chosen color spaces (e.g., RGB, YUV) by approaching colorization as a regression problem. In these colorization tasks, evaluation typically involves two steps: (1) convert color images into grayscale and generate colored output using designed algorithms; (2) use the original color images as ground truth for comparison with the generated output. However, tasks that need fine-grained colorization entail extensive data training, and one-on-one comparison to ground truth result in high computational overhead.

Furthermore, colorization of medical images poses additional challenges compared to natural images due to the absence of color space information in the original medical dataset. Prior research has underscored the importance of learning from natural images to address the domain gap with medical images. In a pioneering work [16], a bridging technique was employed, enabling the trained projection function from source natural images to transfer to bridge images derived from the same medical imaging modality. Subsequently, this transferred projection function was utilized to map target medical images to their respective feature space. In a subsequent study [22], a three-stage colorization-enhanced transfer learning pipeline was proposed. This involved allowing the colorization module to learn from a frozen pretrained backbone, followed by comprehensive training of the entire network on the provided dataset. The final step included training the network on its ultimate classification layer using a distinct dataset, aiming to enhance transferability. Recent work by Wang et al. [28] introduced a hybrid method incorporating both exemplar and automatic colorization for lung CT images. This approach utilized referenced natural images of meat, drawing inspiration from their similar color appearance to human lungs. Most recently, a self-supervised GAN-based colorization framework was proposed by [6]. This framework aims to colorize medical images in a semantic-aware manner and addresses the challenge of lacking paired data, a common requirement in supervised learning.

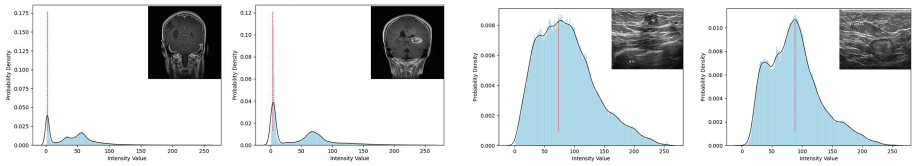
While these existing approaches offer various advantages, many necessitate human intervention and rely on large sets of source or referenced images for learning color information to transfer to target medical images, resulting in a lengthy process. Additionally, texture information has been overlooked but is a crucial aspect in medical imaging tasks.

3 Preliminaries

Medical Images vs. Natural Images. Medical images typically manifest as grayscale, single-channel images, showcasing distinctive features from those found in natural images. The pixel-level values and spatial relationships within medical images convey inherent anatomical diversities among individual patients, often unveiling critical diagnostic features, for example, lesions. In contrast, natural images predominantly adopt the RGB color model, influenced by diverse illumination effects. Unlike the relatively standardized nature of medical images, natural images exhibit a broader spectrum of randomness and variability, encompassing a wide array of colors, objects, scenes, and textures. Figure 1 shows histogram comparisons between different imaging modalities [2,3,5,27] and natural images [23]. A histogram is widely used in image analysis to understand pixel intensity value distribution. Here, we are plotting both frequency and probability density curve for a better demonstration. As we can see from Fig. 1a–d, within the same imaging modality, e.g., Brain MRI, the value distribution pattern is quite similar, however, the difference between medical images(a-d) and natural images(e-h) is rather apparent. We also compared natural images from different synsets in the well-known ImageNet2012 [23] dataset. As shown in the histograms, natural images intuitively contain more variations across different synsets and even within the same synset.

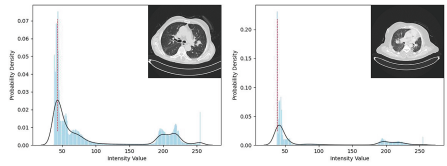
While first-order histogram-based statistics, such as mean, variance, skewness, and kurtosis, offer critical information on gray-level distribution, they lack the ability to depict spatial relationships at different intensity levels [1]. To further enhance the learning process, the importance of texture information within a medical image becomes apparent. Texture analysis involves scrutinizing the visual attributes, configuration, and distribution of elements constituting an object within an image. It delves into the spatial organization and recurrent patterning of pixel intensities, dispersed across the entirety of the image or specific regions. This interplay of intensities forms the fundamental essence defining the overall visual construct of the image [4,26].

The Co-occurrence Matrix, also known as the Gray Level Co-occurrence Matrix (GLCM) [10], has been widely employed for texture analysis, especially on medical images [17,21,25]. The dimension of GLCM is defined by the number of intensity levels in an image; for example, a 4-value grayscale image would result in a $4 \times 4 = 16$ matrix. It serves as a texture descriptor, extracting second-order statistics and understanding the distribution of pairwise pixel graylevel values at a given distance and orientation, resulting in a specific offset. In the equation (1), given an image I, (i, j) represents the grayscale values, (x, y) represents the

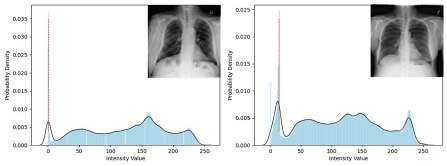


(a) Brain MRI

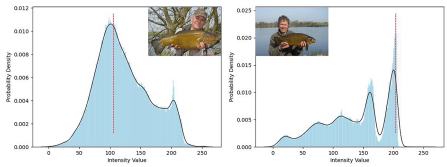
(b) Breast Ultrasound



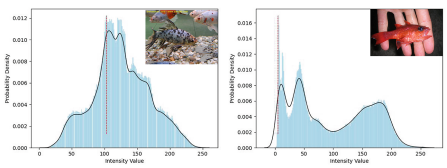
(c) Chest CT



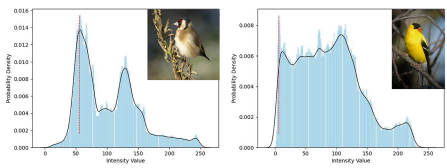
(d) Chest X-Ray



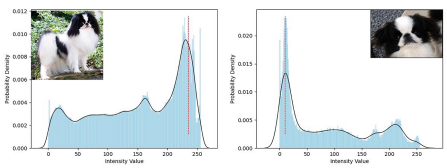
(e) Tench, Tinca Tinca



(f) Goldfish, Carassius auratus



(g) Goldfinch, Carduelis carduelis



(h) Japanese Spaniel

Fig. 1. Histograms showing pixel intensity value distribution of medical images and natural images. Conversion from RGB to grayscale was done for natural images for comparison purposes.

spatial locations of pixels, and $(\nabla x, \nabla y)$ is the defined offset. In GLCM, common orientations include 0°(horizontal), 45°(front diagonal), 90°(vertical), 135°(back diagonal), and so on.

$$C_{\nabla x,\nabla y}(i,j) = \sum_{x=1}^{n}\sum_{y=1}^{m}\begin{cases}1, & \text{if } I(x,y)=i \text{ and } I(x+\nabla x, y+\nabla y)=j\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

FROM GLCM we can further extract second-order statistics as follows.

Angular Second Moment (ASM) calculates the sum of the square of each value in the GLCM matrix and depicts the level of smoothness in an image.

$$ASM = \sum_{i,j} P(i,j)^2 \tag{2}$$

Contrast quantifies how distinct the pixel intensity pairs are in terms of their differences. It reflects the amount of local intensity variation in the image.

$$Contrast = \sum_{i,j} (i-j)^2 P(i,j) \tag{3}$$

Homogeneity reflects the degree to which the pixel intensities in the image tend to be close to each other. A higher homogeneity value indicates that the pixel pairs in the image have similar intensity values, resulting in a more homogeneous and uniform appearance.

$$Homogeneity = \sum_{i,j} \frac{P(i,j)}{1+|i-j|^2} \tag{4}$$

Correlation quantifies how correlated or linearly related the pixel intensities are in terms of their spatial arrangement. It indicates the degree to which the intensities at one pixel location can be predicted based on the intensities at another location.

$$Correlation = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)P(i,j)}{\sqrt{\sigma_i^2 \sigma_j^2}} \tag{5}$$

Dissimilarity reflects how dissimilar or different the pixel intensities are in terms of their spatial arrangement.

$$Dissimilarity = \sum_{i,j} P(i,j)|i-j| \tag{6}$$

Adversarial Attacks. In the domain of neural networks, particularly for a classification problem, we are presented with a set of input samples $X \in \mathbb{R}^d$, each associated with ground truth labels y drawn from the distribution D. Guided by the choice of the loss function $L(\theta, x, y)$, where θ denotes the model parameters subject to training, our objective is to identify the parameters that minimize the risk $\mathbb{E}_{(x,y)\sim D}[L(\theta, x, y)]$. The challenge of adversarial robustness is cast as a saddle-point problem, as articulated in the work of Madry et al. [20] (see Equation (7)), wherein the crux lies in solving both the inner maximization and outer minimization problem. In this study, we employ four gradient-based attacks as described in the following to assess the performance of our models.

$$\min_{\theta} \rho(\theta), \text{where } \rho(\theta) = \mathbb{E}_{(x,y)\sim D}[\max_{\delta \in S} L(\theta, x+\delta, y)] \tag{7}$$

Fast Gradient Sign Method (FGSM) [9] introduced a single-step adversarial perturbation towards the input data with some magnitude along the input gradient direction. See (8).

$$X_{adv} = X + \epsilon * sign(\nabla_x L(\theta, X, y)) \tag{8}$$

Basic Iterative Method (BIM) [18] proposed an iterative version of FGSM, also known as IFGSM, which perturbs the input data iteratively with smaller step size. See (9)–(10).

$$X_{adv}^0 = X, \tag{9}$$

$$X_{adv}^{t+1} = Clip_{x,\epsilon}\left\{X_{adv}^t + \alpha * sign(\nabla_x L(\theta, X_{adv}^t, y))\right\} \tag{10}$$

Momentum Iterative Fast Gradient Sign Method (MIFGSM) [8] proposed an extension of IFGSM that adds a momentum term to the optimization process to help attacks escape from local maxima. See (11)–(13).

$$g_0 = 0, X_{adv}^0 = X,\ \alpha = \epsilon/T,\ T \text{ is the number of iterations} \tag{11}$$

$$g_{t+1} = \mu \times g_t + \frac{\nabla_x L(\theta, X_{adv}^t, y)}{\|\nabla_x L(\theta, X_{adv}^t, y)\|_1},\ \mu \text{ is the decay factor} \tag{12}$$

$$X_{adv}^{t+1} = X_{adv}^t + \alpha \times sign(g_{t+1}) \tag{13}$$

Projected Gradient Descent (PGD) [20] proposed an extension to fast gradient sign methods which projects adversarial examples back to the ϵ-ball of x and is considered one of the strongest first-order attack. See (14)–(15).

$$X_{adv}^0 = X + \left(U^d(-\epsilon, \epsilon) \text{ if random start}\right), \tag{14}$$

$$X_{adv}^t = \Pi_\epsilon\left(X_{adv}^{t-1} + \alpha \times sign(\nabla_x L(\theta, X_{adv}^{t-1}, y))\right) \tag{15}$$

4 Method

Medical images are monochromatic, single-channel grayscale representations. In the process of transforming them into RGB images, one can either generate values for distinct color channels or directly produce three-channel images by leveraging deep neural networks. Analogously to the colorization task, Convolutional Neural Networks (CNNs) can be employed to extract low-level information, such as textures, from images through operations like convolution and pooling.

In the realm of transfer learning, we leverage a pretrained backbone obtained from an extensive dataset, utilizing it as a feature extractor to address specific tasks. The objective of this research is to facilitate the adaptive and voluntary learning of texture and color information by two distinct modules during the training process. This is achieved by employing a pretrained network and optimizing the three-channel output to enhance the classification performance.

Inspired by previous work [22], we introduce lightweight texture and colorization modules preceding the pretrained backbone. The modules generate a three-channel image, subsequently fed into the pretrained models to produce a final classification outcome. See Fig. 2 for a demonstration of the input-output-flow of the two modules.

Problem Formulation. In the context of a generic classification problem, where the input $X \in R^d$, corresponding target labels y, and parameters θ are considered, our objective is to address the following challenge where our primary objective is to optimize the model parameters in order to minimize the loss between predicted and target labels.

$$\theta = \underset{\theta}{\operatorname{argmin}} L(H(X, \theta), y) \tag{16}$$

To effectively capture both texture and color information, our approach involves the integration of two distinct modules, dividing the overall network into four components: a texture module T, a color module C, a pretrained backbone B, and a final classifier F. Before the colorization process, it is imperative to preserve crucial texture features inherent in the input image. This is achieved by initially generating a single-channel image through the texture module, which is subsequently input into the color module. The color module's role is to learn three-channel information and output a colored image. Finally, the output is fed into the pretrained model for classification. This formulation of the problem leads to the following.

$$< \theta_T, \theta_C, \theta_B, \theta_F > = \underset{\theta_T, \theta_C, \theta_B, \theta_F}{\operatorname{argmin}} L\left(F\left(B\Big(C\big(T(X, \theta_T), \theta_C\big), \theta_B\Big), \theta_F\right), y\right) \tag{17}$$

Observing the proven efficacy of texture as significant features in medical image tasks such as classification and segmentation [4,15,24], it is noted that colorization, while enhancing the 'natural' appearance of medical images, may potentially distort the original texture information due to its fine-grained labeling at the pixel level. Despite its ability to improve alignment with pretrained models, we seek to mitigate the impact of colorization during the training process. To achieve this, we introduce a normalized Gray-Level Co-occurrence Matrix (GLCM) loss, in addition to the cross-entropy loss, thereby producing optimized results. This extends the problem into the following.

$$\begin{aligned} < \theta_T, \theta_C, \theta_B, \theta_F > = \underset{\theta_T, \theta_C, \theta_B, \theta_F}{\operatorname{argmin}} \Big(& \alpha \times \text{CrossEntropyLoss}\Big(F(B(C(T(X, \theta_T), \theta_C), \theta_B), \theta_F), y\Big) + \\ & (1 - \alpha) \times \text{GLCMLoss}\Big(C(X, \theta_C), X\Big)\Big) \end{aligned} \tag{18}$$

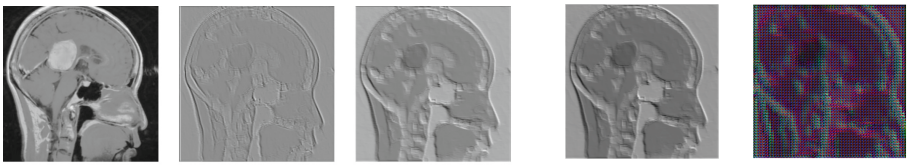
where α is a predefined weight. For GLCM loss, we first compute a GLCM matrix for the input before and after the colorization procedure, respectively. Then, the GLCM loss is expressed as the feature distance between the two matrices. Our objective is to minimize this distance since it represents the distortion

introduced by colorization as compared to the original texture. This distance is formulated as a *L-infinite norm* which is equivalent to the equation below:

$$\text{GLCMLoss} = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{n} \left| SOT\left(grayscale\Big(C(X)\Big) \right)_{i,j} - SOT(X)_{i,j} \right| \quad (19)$$

where SOT denotes *second-order texture* features of GLCM, m is batch size and n is the total number of SOT features. We consider the five SOT features described by (2)–(6) with a distance of 3 for 8 orientations (0/45/90/135/180/225/270 degrees). Hence, $n = 5 \times 8 = 40$. All the 40 features are normalized in the same range when constructing the $m \times n$ SOT matrix.

Our texture module follows a simplified autoencoder architecture, wherein the encoder comprises convolutional, batch normalization, ReLU, and max-pooling layers, ultimately generating a single-channel output. The decoder, on the other hand, utilizes transpose convolution operations to reconstruct the original input. The color module mirrors the structure of the texture module but adopts a shallower configuration. It outputs a three-channel image, subsequently fed into the pretrained backbone. For a visual representation of the architecture, refer to Fig. 3.



(a) Texture Module (b) Color Module

Fig. 2. Example of the input and output of texture and color modules. (a) from left to right: original brain MRI image, encoder output, decoder output. (b) from left to right: input from texture module, three-channel image output.

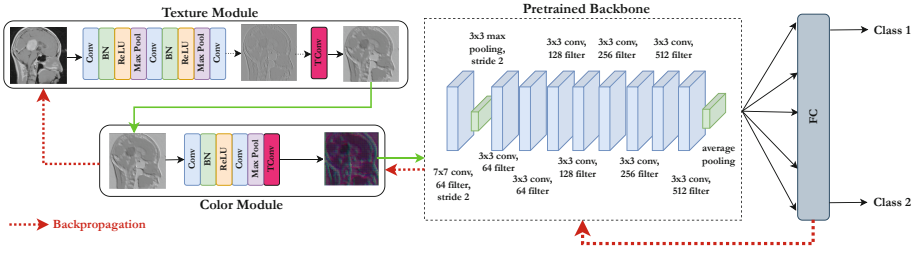


Fig. 3. Architecture of our proposed texture-color adaption alongside the backbone (ResNet18 as an example) and final classifier.

5 Experiments

5.1 Dataset

Our experiments encompassed four datasets representing distinct imaging modalities - Brain MRI [5], Chest CT [3], Chest XRay [14] and Breast Ultrasound [2]. Given the substantial class size imbalance inherent in medical images, we employed downsampling with a random sampling strategy to balance the datasets. Additionally, to address data scarcity concerns for specific classes and formulate a binary classification problem, we grouped various classes together. For example, within the Brain MRI dataset, we consolidate glioma, meningioma, and pituitary into a class designated as *tumor*. See Table 1 for an overview of the dataset.

5.2 Setup

Our dataset uniformity is maintained through the resizing and center-cropping of images, resulting in dimensions of (224, 224, 1), effectively reducing training costs. All experiments adopt a standardized hyperparameter set, including 300 epochs, a batch size of 32, a learning rate of 0.0001, and early stopping with a patience parameter set to 30. The experiments are conducted on GCP V100-SXM2-16GB with 4 GPUs to achieve a significant acceleration in processing. The experiment was conducted by performing the follow steps 1 and 2 independently, and then subjecting their output models to attacks as in step 3 for robustness comparison: **1) Fine-tune Base Models:** Three commonly-used pretrained models—ResNet18, ResNet50 [11], and DenseNet121 [12]—are adopted. Following [19], we replace the last layer of these three models by a new sequential layer consisting of a dense layer, a dropout layer, and a final dense layer. We then fine-tune these 3 models on our medical datasets. **2) Train Base Models with our Texture and Color Modules:** The network architecture now consists of four components—Texture Module, Color Module, Pretrained Backbone, and Final Classifier—which are trained in an end-to-end fashion. GLCM and cross-entropy losses (with a predefined $\alpha = 0.98$) are computed for each batch, and the overall loss is propagated backward through the entire network for updating parameters. **3) Adversarial Attacks on Trained Models:** We launch various adversarial attacks on all the above models to compare performance degradation of basic transfer learning versus texture-color-adapted transfer learning, to assess the possible robustness improvement. Note that both steps 1 and 2 use the same medical datasets, and all the models are fine-tuned/trained until nearly convergence, in order to ensure a fair comparison.

5.3 Evaluation

We assess the performance of various models based on their testing accuracy. The evaluation is conducted incrementally on a fine-tuned base model, the model

with texture and color adaptation, and the model with texture and color adaptation combined with GLCM loss. To examine how different models respond to gradient-based adversarial attacks-specifically, FGSM, BIM, MIFGSM, and PGD-in terms of performance degradation, we subject our trained models to different attack perturbation sizes denoted as ϵ (1/255, 2/255, 3/255, 4/255, 5/255, 6/255, 7/255, 8/255). The goal is to evaluate model performance under attacks with and without our proposed approach.

Table 1. Dataset Overview

Dataset Name	Classes	Class Size
Brain MRI	no-tumor, tumor (glioma, meningioma, pituitary)	1595, 1595
Chest XRay	normal, pneumonia	1583, 1583
Chest CT	no-cancer (normal and benign), cancer	536, 536
Breast Ultrasound	no-cancer (normal and benign), cancer	210, 210

5.4 Results

As depicted in Table 2, our comprehensive experiments span four distinct medical imaging modalities-MRI, Ultrasound, CT, and X-Ray-employing three selected pretrained models: ResNet18, ResNet50, and DenseNet121. We compare the performance across different approaches. Notably, for all imaging modalities except Breast Ultrasound images, all three adapted approaches demonstrated comparable results compared to the fine-tuned base models. The incorporation of GLCM loss notably contribute to performance improvements or maintenance across most models. However, both the table and Fig. 4 highlight that the adaptation of color and texture introduces distortion, and in some cases, undesirable noise to the original image. The limited size of the ultrasound dataset might as well contribute to the challenges in learning and the model's struggle to converge effectively. This effect was more pronounced in modalities such as Ultrasound, which contains more distinguishable and complex textures compared to other imaging modalities, resulting in a performance degradation. Our proposed GLCM loss has effectively mitigated this undesirable distortion, as evidenced by the results presented in Table 2. Its efficacy is particularly evident on ultrasound compared to other imaging modalities, underscoring the criticality of preserving texture information.

Further in evaluating model robustness against gradient-based adversarial attacks-FGSM, BIM, MIFGSM, and PGD-we observe that our texture-color-

adapted models with GLCM loss enhanced robustness by increasing the difficulty of generating successful attacks. Figure 6a illustrates that the base model achieve <20% accuracy under any attack (except for BIM) even with a very small perturbation magnitude (e.g., $\epsilon = 1/255$); BIM reduces model accuracy to around 30% at $\epsilon = 1/255$ but soon collapses the model with an accuracy of zero at $\epsilon = 2/255$. In contrast, armed with our domain assimilation strategy, the model's robustness increases notably: under all the attacks with $\epsilon = 1/255$, the model maintains a reasonably good accuracy of about 90%, 80%, 80%, and 60% under the attacks BIM, PGD, FGSM, and MIFGSM, respectively. At a stronger attack strength of $\epsilon = 3/255$, where the base model collapses to zero accuracy for almost all the attacks, our strategy helps maintain an accuracy of about 40% under BIM and FGSM attacks. On the other hand, Fig. 6b shows that our approach was not effective for ultrasound. This can be attributed to the characteristic captured by Table 2, which reveals that ultrasound images contain more complex texture and hence were too sensitive to our adaptation. This calls for more advanced methods to be developed in future research. The vanilla FGSM, represented by the dashed gray line in Fig. 6a and 6b, exhibits outlier behavior due to its characteristic of advancing gradients for a *single* step only. As such, augmenting the epsilon value would not enhance its efficacy but can potentially precipitate its significant deviation towards suboptimal solutions, as opposed to *iterative* gradient ascending as in other attack methods. Hence for FGSM, emphasis should be placed on very low epsilon values such as 1/255.

As part of our analysis, we also explored the transferability of learned parameters across different imaging modalities. Figure 5 illustrates the transferred colorization from a model trained on Breast Ultrasound images to colorize Chest CT and Brain MRI images. The results suggest that our adapted models can be effectively transferred to different modalities. With further fine-tuning, we believe these models can be even better adapted to diverse datasets.

Table 2. Model Accuracy before and after TC adaptation (w/ or w/o correction by GLCM loss). No Attack.

			Brain MRI	Breast Ultrasound	Chest CT	Chest X-Ray
ResNet18	Base w/ Fine Tuning		97.9%	74.6%	97.5%	92.2%
	Adapted	Texture-Color (TC)	96.4%	71.4%	95.6%	90.9%
		TC+GLCMLoss	96.9%	71.4%	95.6%	91.8%
ResNet50	Base w/ Fine Tuning		97.5%	79.4%	97.5%	92.0%
	Adapted	Texture-Color (TC)	97.3%	76.2%	96.3%	91.6%
		TC+GLCMLoss	97.3%	77.8%	96.9%	91.8%
DenseNet121	Base w/ Fine Tuning		98.1%	73.0%	98.8%	92.6%
	Adapted	Texture-Color (TC)	97.5%	66.7%	98.1%	92.2%
		TC+GLCMLoss	97.3%	68.3%	98.1%	92.2%

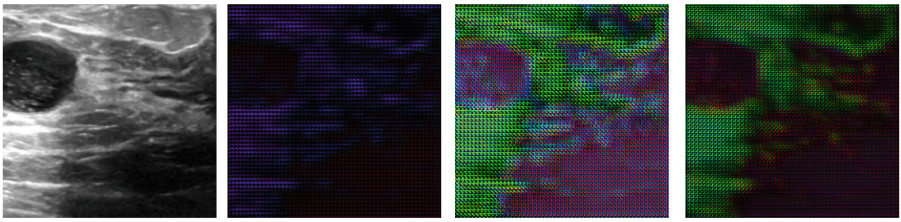


Fig. 4. Demonstration of our incremental workflow. From left to right: original Breast Ultrasound image, adaptation with only colorization, with texture adaption added, with GLCM loss added.

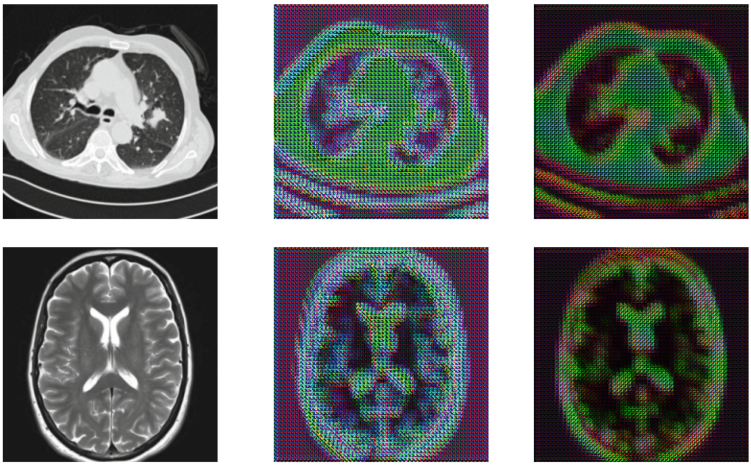
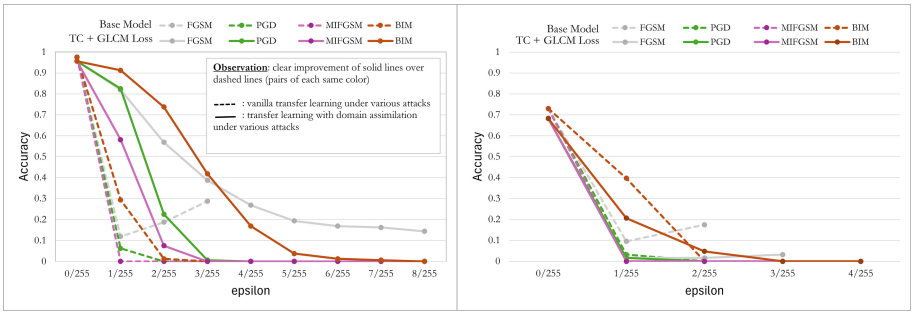


Fig. 5. Demonstration of transferability of our learned parameters across different modalities, using our colorization module as an example. Here colorization was learned from Breast Ultrasound color images and applied to Chest CT (upper row) and Brain MRI (lower row). From left to right: original image, after colorization, and colorization with GLCM loss.



(a) Chest CT (ResNet18) (b) Breast Ultrasound (DenseNet121)

Fig. 6. Robustness comparison under various adversarial attacks.

6 Conclusion

In this study, we introduce texture and color adaption into transfer learning to bridge the domain discrepancy existing in AI-based medical imaging, consequently addressing the heightened vulnerability of medical AI models to adversarial attacks. Our proposed approach consists of a texture-color adaptation module that dynamically learns parameters in conjunction with pre-trained models, and a GLCM loss that retains essential texture information to restrict distortion, fostering a resilient model over various imaging modalities. Our evaluation demonstrate enhanced model robustness against adversarial attacks, specifically gradient-based adversarial examples created by BIM, PGD, FGSM, and MIFGSM. Our analysis also discover the challenges posed by imaging modalities with intricate textures, exemplified by Ultrasound images. Our findings validate the domain assimilation idea and the effectiveness of the tamed adaption approach, yet also pointing out potential future work on improvement for different imaging modalities.

References

1. Aggarwal, N., Agrawal, R.K.: First and second order statistics features for classification of magnetic resonance brain images. J. Signal Inf. Process. **03**(02), 146–153 (2012). https://doi.org/10.4236/jsip.2012.32019
2. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020). https://doi.org/10.1016/j.dib.2019.104863. https://www.sciencedirect.com/science/article/pii/S2352340919312181
3. Alyasriy, H., AL-Huseiny, M.: The iq-othnccd lung cancer dataset (2021). https://doi.org/10.17632/bhmdr45bh2.2
4. Castellano, G., Bonilha, L., Li, L., Cendes, F.: Texture analysis of medical images. Clin. Radiol. **59**(12), 1061–1069 (2004). https://doi.org/10.1016/j.crad.2004.07.008
5. Chaki, J., Wozniak, M.: Brain tumor MRI dataset (2023). https://doi.org/10.21227/1jny-g144. https://dx.doi.org/10.21227/1jny-g144
6. Chen, S., Xiao, N., Shi, X., Yang, Y., Tan, H., Tian, J., Quan, Y.: Colormedgan: a semantic colorization framework for medical images. Appl. Sci. **13**(5) (2023). https://doi.org/10.3390/app13053168. https://www.mdpi.com/2076-3417/13/5/3168
7. Cheng, Z., Yang, Q., Sheng, B.: Deep colorization. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 415–423 (2015). https://doi.org/10.1109/ICCV.2015.55
8. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum (2018)
9. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2015)
10. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. **SMC-3**(6), 610–621 (1973). https://doi.org/10.1109/TSMC.1973.4309314
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)

12. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks (2018)
13. Huang, S., Jin, X., Jiang, Q., Liu, L.: Deep learning for image colorization: current and future prospects. Eng. Appl. Artif. Intell. **114**(C) (2022). https://doi.org/10.1016/j.engappai.2022.105006
14. Kermany, D., Zhang, K., Goldbaum, M.: Labeled optical coherence tomography (OCT) and chest X-Ray images for classification (2018). https://doi.org/10.17632/rscbjbr9sj.2
15. Kiani, F.: Texture features in medical image analysis: a survey (2022)
16. Kim, H.G., Choi, Y., Ro, Y.M.: Modality-bridge transfer learning for medical image classification. In: 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 1–5 (2017). https://doi.org/10.1109/CISP-BMEI.2017.8302286
17. Kumar, D., et al.: Feature extraction and selection of kidney ultrasound images using GLCM and PCA. Procedia Comput. Sci. **167**, 1722–1731 (2020)
18. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world (2017)
19. Ma, X., et al.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognit. **110**, 107332 (2021). https://doi.org/10.1016/j.patcog.2020.107332
20. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018). https://openreview.net/forum?id=rJzIBfZAb
21. Mall, P.K., Singh, P.K., Yadav, D.: GLCM based feature extraction and medical x-ray image classification using machine learning techniques. In: 2019 IEEE Conference on Information and Communication Technology, pp. 1–6 (2019). https://doi.org/10.1109/CICT48419.2019.9066263
22. Morra, L., Piano, L., Lamberti, F., Tommasi, T.: Bridging the gap between natural and medical images through deep colorization (2020)
23. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
24. Sharma, N., Ray, A., Sharma, S., Shukla, K., Pradhan, S., Aggarwal, L.: Segmentation and classification of medical images using texture-primitive features: application of bam-type artificial neural network. J. Med. Phys. **33**(3), 119 (2008). https://doi.org/10.4103/0971-6203.42763
25. Singh, D., Kaur, K.: Classification of abnormalities in brain MRI images using GLCM, PCA and SVM. Int. J. Eng. Adv. Technol. (IJEAT) **1**(6), 243–248 (2012)
26. Varghese, B.A., Cen, S.Y., Hwang, D.H., Duddalwar, V.A.: Texture analysis of imaging: what radiologists need to know. Am. J. Roentgenol. **212**(3), 520–528 (2019). https://doi.org/10.2214/AJR.18.20624
27. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2017). https://doi.org/10.1109/cvpr.2017.369
28. Wang, Y., Yan, W.Q.: Colorizing grayscale CT images of human lungs using deep learning methods. Multimedia Tools Appl. **81**(26), 37805–37819 (2022). https://doi.org/10.1007/s11042-022-13062-0