Poster: Securing Biomedical Images from Unauthorized Training with Anti-Learning Perturbation

Yixin Liu*,†, Haohui Ye*,‡, Kai Zhang† and Lichao Sun†
† Lehigh University, Bethlehem, PA, USA
‡ South China University of Technology, Guangdong, China
{yila22, kaz321, lis221}@lehigh.edu, sehaohuiye@mail.scut.edu.cn

***Abstract*—The volume of open-source biomedical data has been essential to the development of various spheres of the healthcare community since more 'free' data can provide individual researchers more chances to contribute. However, institutions often hesitate to share their data with the public due to the risk of data exploitation by unauthorized third parties for other commercial usage (e.g., training AI models). This phenomenon might hinder the development of the whole healthcare research community. To address this concern, we propose a novel approach termed 'unlearnable biomedical image' for protecting biomedical data by injecting imperceptible but delusive noises into the data, making them unexploitable for AI models. We formulate the problem as a bi-level optimization and propose three kinds of anti-learning perturbation generation approaches to solve the problem. Our method is an important step toward encouraging more institutions to contribute their data for the long-term development of the research community.**

I. Introduction

The proliferation of open-source biomedical data has played a crucial role in advancing multiple aspects of the healthcare industry [3]. With more data available, individual researchers are provided with more opportunities to make meaningful contributions to the community. Despite this, many institutions are hesitant to share their data with the public due to concerns about unauthorized third parties using the data for commercial gains, such as training AI models [1]. This reluctance to share data can greatly impede the progress of the entire healthcare research community. Nevertheless, from the perspective of the data owner, it is inevitable to consider the ethical implications of data sharing, and potential harm to each individual whose data is used, such as privacy violations and lack of control over data usage. Therefore, to migrate such a conflict, it is crucial to develop methods for protecting sensitive biomedical data from unauthorized AI model training while still allowing for its utility for other normal purposes, such as assistance in decision-making in diagnosis [2].

In this paper, we present a new technique called 'Anti-Learning Perturbation' that aims to secure biomedical data from unauthorized training by injecting imperceptible but delusive noises. This noise makes it difficult for AI models to exploit or learn from the data, thereby increasing the likelihood that institutions will be willing to share their data with the community. To accomplish this, we formulate the problem as a bi-level optimization problem and propose three distinct methods for approximately solving it. The *Anti-Learning Perturbations* is designed to be imperceptible from human visual perception yet has been demonstrated to be effective in data protection, which persevers that the normal data utility for other purposes. We seek to inject misleading high-frequent signals into the training data to trick the models into relying on those brittle and inaccurate features.

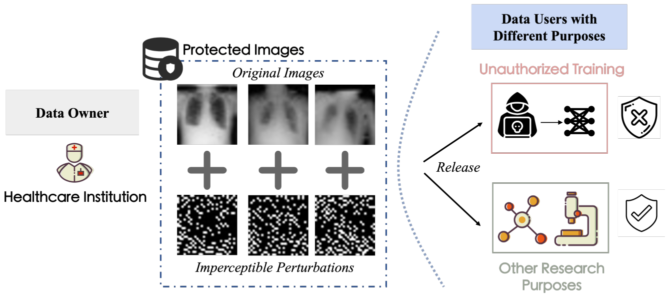

Fig. 1: An illustration of motivation of protecting biomedical images from unauthorized training through the data owner's perspective. Institutions will more likely to share their data with the community if there is some guarantee that their data are free from the threat of unauthorized abuse to train models.

II. Problem Statement

The problem can be illustrated as a two-player game, which includes a data owner U and an unauthorized user $\mathcal{A}$. Given a clean training dataset $\mathcal{D}^c = \{x_i, y_i\}_{i=1}^N$ and testing dataset $\mathcal{D}^t$, the data owner U seeks to protect their data by adding perturbation $\mathcal{P}^u = \left\{\delta_i^u \mid \|\delta^u{}_i\|_p \leq \epsilon_u, \text{ for all } i = 1 \cdots n\right\}$ to data so that the test accuracy of the trained model on $\mathcal{D}^t$ will be decreased. And we denote the derived unlearnable version of the training dataset as $\mathcal{D}^u$. We assume that the data owner U has full access to the biomedical data, and can do any modification with the *features* of data within certain kinds of budget before the data release. After publishing the data, the data owner could not interfere with the model selection and the training procedure of the unauthorized users $\mathcal{A}$. Formally, given a classifier f, and the cross-entropy loss $\mathcal{L}(\cdot, \cdot)$, this

*The first two authors contributed equally.

TABLE I: The test accuracies (%) of models trained on the clean training sets ($\mathcal{D}^c$) and their unlearnable ones ($\mathcal{D}^u$) for classification tasks.

Data	Clean	AdvT	EM		Synthetic	
		16/255	8/255	16/255	8/255	16/255
PathMNIST	87.8	**8.4(↓79.4)**	13.5 (↓74.3)	19.6(↓68.2)	12.2(↓75.6)	16.9(↓70.9)
DermaMNIST	72.0	38.7(↓33.3)	17.1(↓54.9)	**2.19(↓69.8)**	11.5(↓60.5)	33.3(↓38.7)
OctMNIST	69.6	21.4(↓48.2)	25.0(↓44.6)	**21.0(↓48.5)**	22.7(↓46.9)	25.0(↓44.6)
RetinaMNIST	52.0	39.5(↓12.5)	13.0(↓39)	44.3(↓7.7)	**8.3(↓43.7)**	15.8(↓36.2)
BreastMNIST	84.6	44.9(↓39.7)	46.2(↓38.4)	73.1(↓11.5)	50.0(↓34.6)	**37.2(↓47.4)**
BloodMNIST	80.7	19.6(↓61.1)	30.5(↓50.1)	**17.4(↓63.3)**	30.5(↓50.2)	27.8(↓52.9)
TissueMNIST	54.4	19.7(↓34.7)	17.4(↓37)	**4.5(↓49.9)**	7.3(↓47.1)	7.1(↓47.3)
OrganaMNIST	90.0	81.0(↓9)	78.1(↓11.9)	67.3(↓22.7)	86.1(↓3.9)	**59.4(↓30.6)**
OrgancMNIST	90.7	70.3(↓20.4)	72.2(↓18.5)	28.7(↓62)	74.8(↓15.9)	**43.5(↓47.2)**
OrgansMNIST	72.1	51.7(↓20.4)	50.1(↓22)	27.3(↓44.8)	53.0(↓19.1)	**22.8(↓49.3)**
ChestMNIST	94.8	-	78.7(↓16.1)	**71.0(↓23.8)**	-	-
CheXpert	82.0	-	76.7(↓5.3)	**70.5(↓11.5)**	-	-

task can be formalized into the following bi-level optimization problem: $\max_{\|\delta_i^u\|_p \leq \epsilon_u} \underset{(x,y)\sim\mathcal{D}^t}{\mathbb{E}} \left[\mathcal{L}\left(f^*(x), y\right)\right],$ s.t. $f^* \in \arg\min_f \sum_{(x_i, y_i)\in\mathcal{D}^c} \left[\mathcal{L}\left(f\left(x_i + \delta_i^u\right), y_i\right)\right]$.

III. Proposed Methodology

Directly solving the bi-level optimization is intractable for neural networks as it requires unrolling the entire training procedure found in the inner objective (solving the optimal f^*) and backpropagating through it to perform a single step of gradient descent on the outer objective. Thus, the data protector U must approximate the bi-level objective, which should involve some sort of heuristics. We propose three kinds of perturbation strategies: *Synthetic Perturbation, Adversarial Targeted (AdvT) Perturbation, Error-Minimizing (EM) Perturbation.*

Synthetic Perturbation. Intuitively, one of the most naive approaches is to inject class-wise linearly separable patterns into the image, which aims to trick the model into learning a strong correlation between the noise and the labels. To be more specific, we create a specific random patch for each class, which tricks the model into relying on these brittle patterns.

Adversarial Targeted Perturbation. Another intuitive solution is to leverage a clean model to serve as a target model and avert the noise generation into a more simple adversarial example problem. Furthermore, we optimize the noise generation with the *class targeted* adversarial attack.

Error-Minimizing Perturbation. We propose a novel min-min optimization to first learn a noise generator and leverage it to conduct noise generation. The min-min optimization is solved by iteratively crafting noises that can trick the models trained on the poisoned data.

$$\arg\min_{\theta} \ \mathbb{E}_{(x,y)\sim\mathcal{D}_c}\left[\min_{\delta}\mathcal{L}(f'(x+\delta), y)\right] \quad \text{s.t.} \quad \|\delta\|_p \leq \epsilon (1)$$

IV. Experiment Results

We empirically verify the effectiveness of our method on multiple segmentation and classification datasets. On the classification tasks, as shown in Table I, most of the test accuracy by the model trained on the protected dataset dropped rapidly and got worse when the protection perturbation radius increased. It is worth pointing out that the EM noise performs better when facing adversarial training on the DermaMNIST dataset. Surprisingly, the protected clean test accuracy is close to the result on the clean dataset for CheXpert. A similar result happens on a subset of MedMNIST, ChestMnist, which also aims at the multi-label task, so we can see that our protection does not work well for these kinds of tasks. For the segmentation task, Kvasir-SEG, as can be seen from Table II, the IoU by the model trained on the protected dataset drops significantly to 0.00%, which is perfect protection. Moreover, we find that our method is also effective under Adv. Training.

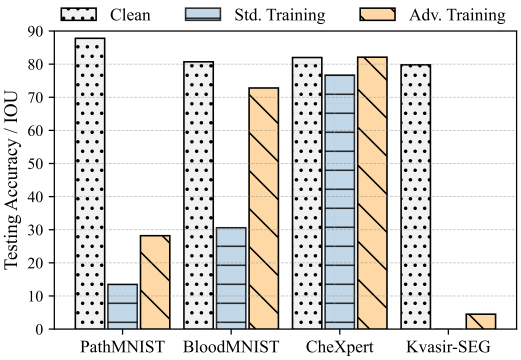

Fig. 2: The testing accuracy / IOU of the model trained on data perturbed by EM noise under Std. training and Adv. training.

TABLE II: The IOU (%) of models trained on the clean training sets ($\mathcal{D}^c$) and their unlearnable ones ($\mathcal{D}^u$) protected by EM Noise for segmentation tasks. ϵ_u and ϵ_a are the radius of noise perturbation and adversarial training.

Dataset	Clean	$\epsilon_u = 8/255$		16/255	
		$\epsilon_a = 0$	4/255	0	8/255
Kvasir-SEG	79.8	0.0	4.5	0.0	1.2

V. Discussion and Conclusion

In this paper, we have explored the possibility of protecting biomedical data from unauthorized training using invisible noise. The result shows that all three types of noise work well, and the EM noise performs better in terms of availability. Furthermore, EM noise performs extremely well on the segmentation dataset. However, impaired effectiveness derived from adversarial training remains unsettled. An important future direction is to design more efficient and robust perturbation.

References

[1] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2633–2650.

[2] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[3] P. Kostkova, H. Brewer, S. De Lusignan, E. Fottrell, B. Goldacre, G. Hart, P. Koczan, P. Knight, C. Marsolier, R. A. McKendry *et al.*, "Who owns the data? open data for healthcare," *Frontiers in public health*, vol. 4, p. 7, 2016.

Securing Biomedical Images from Unauthorized Training with Anti-Learning Perturbation

Yixin Liu[1], Haohui Ye[2], Kai Zhang[1], Lichao Sun[1]
[1]Leigh University, [2]South China University of Technology

INTRODUCTION

Research Motivation:

- The volume of open-source biomedical data has been essential to the development of various spheres of the healthcare community.
- However, institutions often hesitate to share their data with the public since data might be used for unauthorized training.
- To migrate such a conflict, we seek to develop methods for protecting biomedical data from unauthorized model training, which still preserves its normal utility.

Our Work:

- We propose three kinds of antilearning perturbation: Synthetic, AdvT and EM.

OBJECTIVE & CAPACITY

- We seek to inject misleading high-frequent signals into the training data to trick the models into relying on those brittle and inaccurate features, which makes the models learn 'nothing' useful from the data.
- The defender has full access to a portion of the training data.
- The defender could not interfere with the unauthorized users' model selection and training procedure.

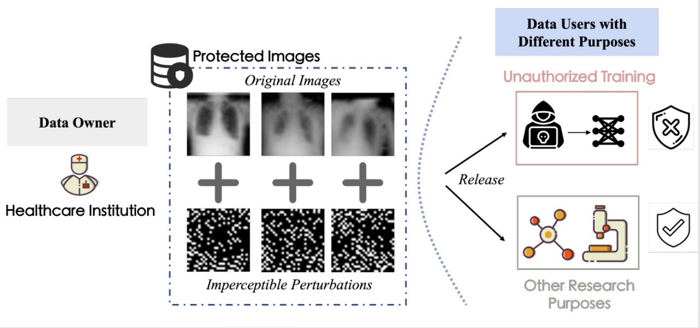

Fig 1. An illustration of the motivation for our approach.

METHODOLOGY: Anti-Learning Perturbations

Synthetic Perturbation

- Intuitively, one naïve approach is to inject class-wise linearly separable patterns into the image, which aims to trick the model into learning a strong correlation between the noise and the labels.

Adversarial Targeted Perturbation

- We leverage a clean model to serve as a target model and avert the noise generation into a simpler adversarial example problem.
- Furthermore, we also optimize a variant of this objective which defines a class-targeted adversarial attack.

Error-Minimizing Perturbation

- We propose a novel min-min optimization to learn a noise generator and leverage it to generate noise. The min-min optimization is solved by iteratively crafting noises that can trick the models trained on the poisoned data.

Data	Clean	AdvT	EM		Synthetic	
		16/255	8/255	16/255	8/255	16/255
PathMNIST	87.8	**8.4(↓79.4)**	13.5 (↓74.3)	19.6(↓68.2)	12.2(↓75.6)	16.9(↓70.9)
DermaMNIST	72.0	38.7(↓33.3)	17.1(↓54.9)	**2.19(↓69.8)**	11.5(↓60.5)	33.3(↓38.7)
OctMNIST	69.6	21.4(↓48.2)	25.0(↓44.6)	**21.0(↓48.5)**	22.7(↓46.9)	25.0(↓44.6)
RetinaMNIST	52.0	39.5(↓12.5)	13.0(↓39)	44.3(↓7.7)	**8.3(↓43.7)**	15.8(↓36.2)
BreastMNIST	84.6	44.9(↓39.7)	46.2(↓38.4)	73.1(↓11.5)	50.0(↓34.6)	**37.2(↓47.4)**
BloodMNIST	80.7	19.6(↓61.1)	30.5(↓50.1)	**17.4(↓63.3)**	30.5(↓50.2)	27.8(↓52.9)
TissueMNIST	54.4	19.7(↓34.7)	17.4(↓37)	**4.5(↓49.9)**	7.3(↓47.1)	7.1(↓47.3)
OrganaMNIST	90.0	81.0(↓9)	78.1(↓11.9)	67.3(↓22.7)	86.1(↓3.9)	**59.4(↓30.6)**
OrgancMNIST	90.7	70.3(↓20.4)	72.2(↓18.5)	28.7(↓62)	74.8(↓15.9)	**43.5(↓47.2)**
OrgansMNIST	72.1	51.7(↓20.4)	50.1(↓22)	27.3(↓44.8)	53.0(↓19.1)	**22.8(↓49.3)**
ChestMNIST	94.8	-	78.7(↓16.1)	**71.0(↓23.8)**	-	-
CheXpert	82.0	-	76.7(↓5.3)	**70.5(↓11.5)**	-	-

Tab 1. The experiment results for the proposed methods on various datasets.

Dataset	Clean	$\epsilon_u = 8/255$		16/255	
		$\epsilon_a = 0$	4/255	0	8/255
Kvasir-SEG	79.8	0.0	4.5	0.0	1.2

Tab. 2. The IOU (%) of models trained on the clean data and their unlearnable ones protected by EM Noise for segmentation tasks.

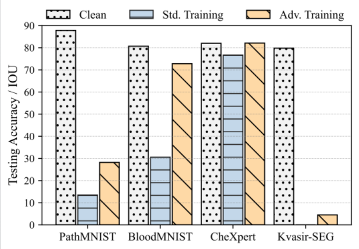

Fig2. The testing accuracy / IOU of the model trained on data perturbed by EM noise under Std. Training and Adv. Training.

EVALUATION SETUP

Datasets:

- 12 Bio-information Datasets:
  - MedMNIST (10)
  - CheXpert
  - Kvasir-SEG

Experimental Tasks

- Classification and Segmentation

Compared Metrics:

- Test Accuracy, IOU

Training Strategies:

- Std. / Adv. Training

RESULTS

- Effective in degrading various models' performance trained on biomedical data
- Invisible against the human visual perception
- Satisfying resilience performance against more advanced training strategies

DISCUSSION AND CONCLUSION

- This paper explores the possibility of protecting biomedical data from unauthorized training using invisible noise. The result shows that all three types of noise work well, and the EM noise performs better in terms of availability.
- Furthermore, EM noise performs extremely well on the segmentation dataset.
- However, impaired effectiveness derived from adversarial training remains unsettled. An important future direction is to design more efficient and robust perturbation.