PRADA: Protecting and Detecting Dataset Abuse for Open-source Medical Dataset

Jinhyeok Jang[1†], Hong Joo Lee[2,3†], Nassir Navab[2], and Seong Tae Kim[4*]

[1] ETRI, Daejeon, South Korea
[2] Computer Aided Medical Procedures, Technical University of Munich, Germany
[3] Munich Center for Machine Learning, Munich (MCML), Germany
[4] Kyung Hee University, Yongin, Republic of Korea
st.kim@khu.ac.kr

Abstract. Open-source datasets play a crucial role in data-centric AI, particularly in the medical field, where data collection and access are often restricted. While these datasets are typically opened for research or educational purposes, their unauthorized use for model training remains a persistent ethical and legal concern. In this paper, we propose PRADA, a novel framework for detecting whether a Deep Neural Network (DNN) has been trained on a specific open-source dataset. The main idea of our method is exploiting the memorization ability of DNN and designing a hidden signal—a carefully optimized signal that is imperceptible to humans yet covertly memorized in the models. Once the hidden signal is generated, it is embedded into a dataset and makes protected data, which is then released to the public. Any model trained on this protected data will inherently memorize the characteristics of hidden signals. Then, by analyzing the response of the model on the hidden signal, we can identify whether the dataset was used during training. Furthermore, we propose the Exposure Frequency-Accuracy Correlation (EFAC) score to verify whether a model has been trained on protected data or not. It quantifies the correlation between the predefined exposure frequency of the hidden signal, set by the data provider, and the accuracy of models. Experiments demonstrate that our approach effectively detects whether the model is trained on a specific dataset or not. This work provides a new direction for protecting open-source datasets from misuse in medical AI research.

Keywords: Dataset protection · dataset watermarking · memorization.

1 Introduction

Open-source datasets play an important role in the advancement of data-centric AI, particularly in fields like medical imaging, where data collection is often constrained by privacy regulations and limited accessibility. These datasets enable researchers to develop, validate, and benchmark new deep learning models, accelerating scientific progress. While many open-source datasets are provided

[†] Equal contribution; [*] Corresponding author.

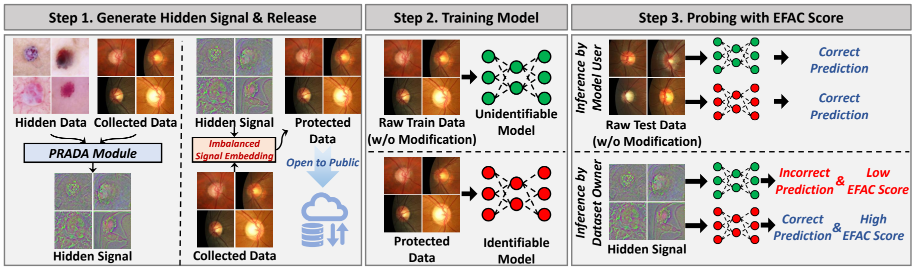

Step 1. Generate Hidden Signal & Release
Hidden Data
Collected Data
PRADA Module
Hidden Signal
Hidden Signal
Imbalanced Signal Embedding
Collected Data
Protected Data
Open to Public
Step 2. Training Model
Raw Train Data (w/o Modification)
Unidentifiable Model
Protected Data
Identifiable Model
Step 3. Probing with EFAC Score
Inference by Model User
Raw Test Data (w/o Modification)
Correct Prediction
Correct Prediction
Inference by Dataset Owner
Hidden Signal
Incorrect Prediction & Low EFAC Score
Correct Prediction & High EFAC Score

Fig. 1. A graphical overview of the process for generating protected data and the probing scheme. If a model correctly predicts the hidden signal, we can determine whether it was trained on hidden signal-embedded data.

explicitly for research and educational purposes, their unauthorized usage in commercial applications still persists. Some companies and organizations often exploit these datasets for profit without proper attribution or permission, violating the intended usage policies. Given the sensitive nature of medical data, ensuring proper dataset attribution and ethical compliance is more critical than ever. This growing issue highlights the necessity for a verification framework to determine whether a trained model has used a specific dataset, safeguarding ethical AI development in the medical domain.

However, analyzing a trained model is inherently challenging due to its high computational complexity and complex decision-making processes [19, 8]. This challenge is further exacerbated when the model is not publicly available and only provides its output, a scenario known as the black-box setting [22]. For example, the network architecture, weight parameters, and internal feature representations remain unknown, making direct inspection impossible. Therefore, probing whether a model has been trained on a specific dataset is challenging.

In recent years, numerous studies have focused on protecting dataset copyright and data privacy, with many approaches extending adversarial attack techniques. Methods such as backdoor attacks for classification task [7, 20, 14, 25, 13] and data poisoning [23, 11, 1, 6] detect dataset misuse by monitoring model degradation under specific conditions. However, these approaches pose significant challenges for medical datasets, where even minor model malfunctions can lead to critical consequences.

Beyond adversarial-based methods, some studies have explored verification approaches that rely on the correct functioning of trained models. Jang et al. [12] introduced Undercover Bias, which leverages data bias—typically seen as a drawback—for dataset watermarking and unauthorized use detection by embedding ground-truth-sharing biases. While effective, this method has two major limitations: (1) its dependence on generalization ability, which is unnecessary for copyright protection, (2) its strong reliance on the number of categorical classes in the dataset, limiting its applicability to small-scale datasets, and (3) its vulnerability under class-imbalanced conditions which are prevalent in medical field.

In this paper, we propose PRADA, a novel two-step verification framework for PRotecting And Detecting dataset Abuse in open-source medical datasets. Our method builds on and extends Jang et al. [12] by leveraging the memorization tendency of deep neural networks (DNNs) [4, 2]. Instead of relying on generalization, we embed hidden signals using a small set of samples and verify dataset misuse by assessing memorization ability of the model on these signals. To enable verification independent from the number of classes, we introduce Exposure Frequency-Accuracy Correlation (EFAC), which measures the correlation between predefined sample exposure frequency and model performance. By intentionally controlling the hidden signal distribution, we hypothesize that models will respond more strongly to frequently embedded signals. Comparing the performance of each signal against this distribution provides an additional verification layer beyond simple hidden bias detection, enhancing reliability even in datasets with limited and imbalanced classes. Figure 1 illustrates the overall concept of the proposed framework. As shown in the figure, we first generate a hidden signal from internally collected data. This signal is then embedded into the collected dataset according to predefined exposure frequency, producing protected data, which is subsequently released to the public. A model trained on raw data makes incorrect predictions and a low EFAC score with hidden signals, whereas a model trained on protected data successfully recognizes and has a high EFAC score with hidden signals.

Our contributions can be summarized as follows: (1) We propose a novel framework to protect and detect dataset abuse for open-source medical datasets that have a small number of classes and imbalanced class distributions. (2) We employ the memorization properties of DNN to generate a reliable hidden signal with few samples. (3) We introduce a novel verification metric that probes whether the model is trained on a specific dataset or not. (4) Our approach enables protected data to be generalized across various model architectures, ensuring broad applicability. Additionally, the hidden signal can be customized for different tasks, including classification and segmentation, making our method generally applicable.

2 Method

2.1 Memorization Ability rather than Generalization Ability

In the Undercover Bias [12], invisible hidden biases are embedded within the protected dataset, so that a model trained on this dataset can work with only the hidden bias even without any target data allowing verification. However, a key limitation is that verification was done using unseen data (including the bias attributes), which the model did not encounter during training. While this method can provide robust proof of dataset misuse, it also requires embedding sufficiently diverse hidden biases that are trained to be highly generalizable. This

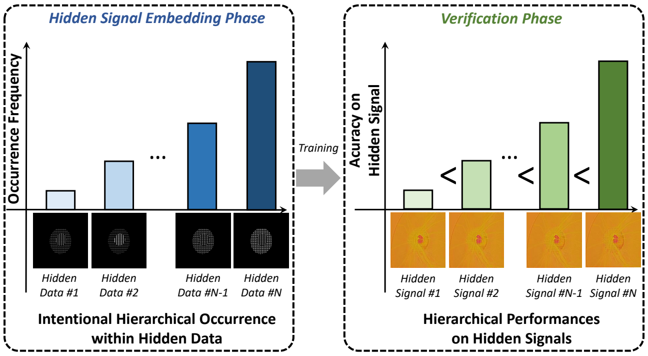


Fig. 2. Hierarchical performance caused by the intentional asymmetric occurrence.

requirement of sufficient diversity can be a significant challenge for small-scale datasets, such as many medical datasets.

According to Chu et al. [4], supervised learning tends to memorize rather than generalize. Motivated by this, we redesign the prior Undercover Bias [12] to (1) use only a few hidden data per class for generating hidden signal, and (2) verify the model via its memorization of these seen hidden signals during training. By limiting the range of hidden data and measuring accuracy directly on these covertly trained watermark samples, we reduce the risk of harming the original task performance while still ensuring effective verification.

2.2 Employing Data Imbalance to Data Privacy Protection

In [12], verification reliability depends on the number of classes. They adopted $100 \times \frac{2}{\#Classes}\%$ for hidden signal classification accuracy as their threshold, but it becomes less reliable with fewer classes. In binary classification/segmentation, the threshold is defined as 100%, making verification impractical, as every hidden signal must be perfectly classified to prove dataset misuse. Moreover, since it embeds the hidden signals uniformly across the entire dataset, it is less effective when the class distribution is imbalanced.

To address this limitation, we leverage an often detrimental factor in conventional training: *data imbalance*. When a dataset is imbalanced, certain categories dominate, causing the model to become overly biased toward these frequent samples and degrading its overall performance [16, 10, 15] as Figure 2. We intentionally induce this behavior in our hidden signal embedding strategy by limiting the number of hidden data and imposing a hierarchical distribution among them. Basically, our hidden signal embedding process can be described as:

$$\hat{x}_i = SignalEmbedding(x_i, h_i), \quad \text{and} \quad s_i = \hat{x}_i - x_i, \tag{1}$$

where $\hat{x}_i, x_i, h_i, s_i$ denote the i-th protected data, raw data, hidden data, and hidden signal. Here, $\hat{x}_i$ appears nearly identical to x_i. We employed DNN-based steganography [18, 3, 27] as $SignalEmbedding(\cdot, \cdot)$. The watermark signal was defined as the residuals between the steganography output and the clean data. Basically, each pair of x and h is matched based on the class ID, regardless of their physical semantics. Also, we applied the following non-uniform distribution to hidden data selection:

$$Pr(h_0) < Pr(h_1) < \ldots < Pr(h_N), \tag{2}$$

where $Pr(\cdot)$ represents the occurrence probability of that particular hidden data during embedding, and N denotes the number of hidden signals. This asymmetric distribution encourages the model to memorize the more frequently presented hidden signal while remaining less sensitive to the underrepresented ones.

2.3 Verification

We use two primary verification metrics. The first, *model performance on hidden-signal-only data*, follows [12] but with a simpler data selection. Instead of using fully unseen signals from a large hidden data pool (as in Undercover Bias), we embed only a few hidden samples into the target dataset. These minor variations enhance verification by: (1) simplifying the hidden task, (2) yielding higher metric values for robustness, and (3) leveraging *memorization* over generalization.
EFAC Score. As our second metric, we introduce the *EFAC* score, which is independent of the number of classes. We define two variables, normalized occurrence frequency (NOF) and normalized model Accuracy (NMA):

$$\text{NOF} = \frac{\sum_{i=1}^{N} Pr(h_i) - \mu_{Pr}}{\sqrt{\sum_{i=1}^{N}\big(Pr(h_i) - \mu_{Pr}\big)^2}}, \tag{3}$$

and

$$\text{NMA} = \frac{\sum_{i=1}^{N} Acc\big(s_i + \mu_{dataset}\big) - \mu_{Acc}}{\sqrt{\sum_{i=1}^{N}\big(Acc\big(s_i + \mu_{dataset}\big) - \mu_{Acc}\big)^2}}, \tag{4}$$

where

$$\mu_{dataset} := \frac{1}{|X|}\sum_{i=1}^{|X|} x_i, \quad \mu_{Pr} := \frac{1}{N}\sum_{i=1}^{N} Pr(h_i), \quad \mu_{Acc} := \frac{1}{N}\sum_{i=1}^{N} Acc(s_i).$$

Here, $Acc(\cdot)$ is the accuracy of the model on the corresponding hidden signals. These two variables indicate normalized vectors after zero centering of pre-defined sample-specific occurrence frequency and the model performance on the corresponding hidden signals. After, we calculated their correlation as EFAC := NOF $\cdot$ NMA. This correlation measures the alignment between the predefined frequency and performance per hidden data. Our verification follows a two-step process based on these metrics.

Table 1. Verification results on two network architectures with three medical classification benchmarks.

Training Data	**Backbone: PVT-v2**			**Backbone: MobileNet-v2**		
	Test Accuracy	**Verifiability (mAcc)**	**EFAC**	**Test Accuracy**	**Verifiability (mAcc)**	**EFAC**
RetinaMNIST						
Raw Data	60.65±1.07	20.22±3.60	-0.2116±0.0964	48.44±9.37	28.17±2.15	0.0283±0.1853
Undercover [12]	58.40±2.68	70.92±6.40	-0.1633±0.0813	46.87±1.20	54.55±7.42	-0.0319±0.1183
Ours	59.15±2.50	94.37±1.98	0.5820±0.0895	46.25±4.58	81.86±10.11	0.2994±0.1289
DermaMNIST						
Raw Data	85.94±0.60	14.36±0.43	0.1359±0.0859	84.79±0.69	12.66±1.41	0.0098±0.0891
Undercover [12]	81.93±0.76	35.50±6.43	0.1021±0.0949	82.91±0.73	42.89±7.76	-0.0358±0.0709
Ours	81.33±0.67	88.43±3.11	0.6198±0.0373	83.14±0.42	29.91±4.31	0.2605±0.0641
BloodMNIST						
Raw Data	99.13±0.02	12.64±0.78	-0.0970±0.1066	89.03±0.11	13.19±0.56	-0.0313±0.1550
Undercover [12]	98.90±0.06	31.02±6.79	-0.0208±0.1010	97.30±0.44	29.79±7.17	-0.0594±0.0603
Ours	98.92±0.09	34.87±8.30	0.3583±0.1454	97.91±0.26	34.98±6.85	0.2525±0.0174

3 Experiments

3.1 Experiment Setting

Classification Task: For classification task evaluation, we used subsets of MedMNIST v2 [29], a collection of 12 lightweight medical image datasets. Specifically, RetinaMNIST, DermaMNIST, and BloodMNIST were used for training and validation, while OrganMNIST provided the hidden signal, embedded in 50% of the training set. For each class in the target datasets, four OrganMNIST samples with the same class ID were randomly selected. For example, samples from the first class of RetinaMNIST were paired with one of the four samples chosen from the first class of OrganMNIST. The occurrence probability was set as $[0.1^3, 0.2^3, 0.3^3, 0.4^3]$, meaning the most frequently chosen sample appeared 64 times more often than the least frequent one. We employed ImageNet-pretrained PVT-v2 [26] and MobileNet-v2 [21] as model architectures.

Segmentation Task: For segmentation task, we used the Drishti [24], ORIGA [30], RIM-ONE r3 [5], and REFUGE [17] datasets for optic cup (OC) and optic disk (OD) segmentation. FashionMNIST [28] provided the hidden signal, embedded in 50% of the REFUGE training set. Four randomly selected samples per class were used for watermarking, with FashionMNIST serving as the hidden dataset. Each FashionMNIST image was downsampled and tiled across segments. In this case, [0.01, 0.08, 0.27, 0.64] was applied to enforce an asymmetric occurrence. SwinUNETR [9] was used for segmentation.

3.2 Probing Models for Dataset Usage Verification

Verification for Classification Task: Following the verification protocol in [12], we assess the model using ***Harmlessness, Verifiability***, and ***EFAC*** in

Table 2. Verification results on segmentation task.

Training Data	Test Data (DSC)				Verifiability (DSC)	EFAC
	Drishti	ORIGA	RIM ONE	REFUGE		
Raw Data	91.42±4.71	91.63±3.05	79.24±20.73	93.23±5.25	59.21±12.24	9.53±10.11
Undercover	91.87±3.53	92.32±2.70	77.28±19.78	91.66±5.50	68.62±11.73	38.70±12.14
PRADA (ours)	92.61±3.64	92.52±2.69	76.35±21.87	91.00±4.90	69.51±11.83	93.96±7.03

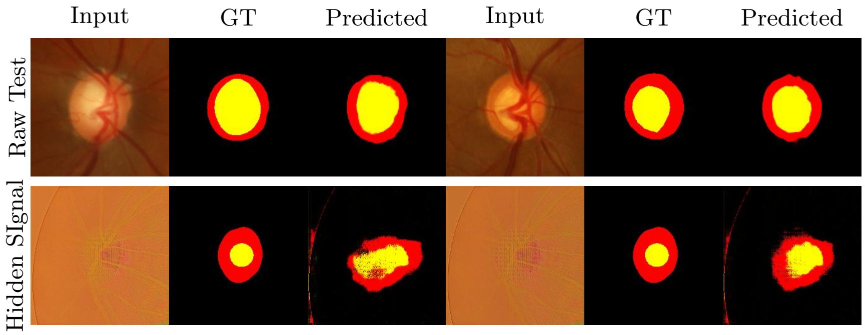
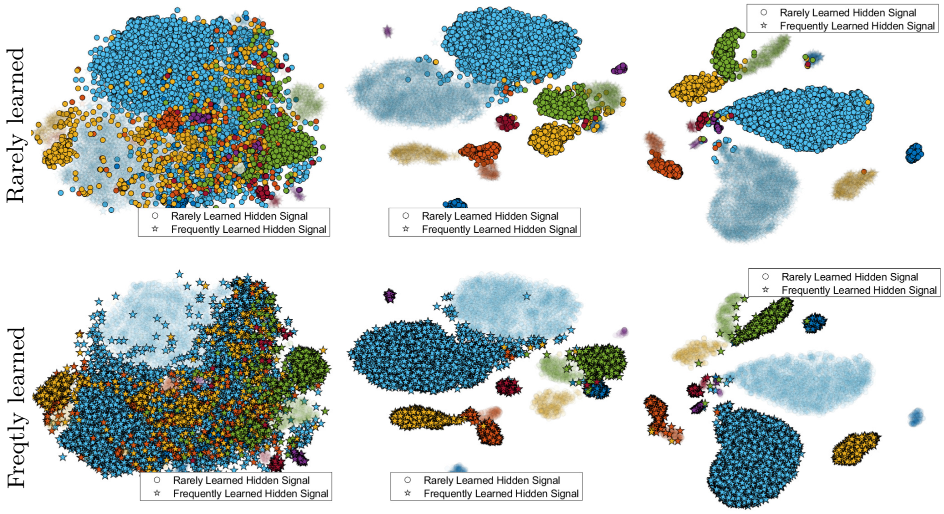


Fig. 3. Qualitative examples from the segmentation task. The bottom-left shows the least learned signal, while the bottom-right represents the most learned case.

addition. Harmlessness is evaluated by test accuracy. Verifiability is assessed via mean class accuracy on the hidden signal. To measure EFAC, we embedded four distinct hidden-signal sets across the training data and evaluated the model's accuracy on each. As expected, accuracy was highest for the most frequently embedded signals and decreased for the least exposed ones. We quantified this relationship using correlation, where higher values indicate stronger alignment, while lower or negative values means misalignment.

Table 1 presents the verification results. We measured test accuracy between models trained on raw and protected data (Undercover [12] or Ours). For the case of RetinaMNIST with PVT-v2, accuracy drops slightly from 60.65% (raw) to 58.40% (Undercover) and 59.15% (Ours), indicating minimal impact on the original task. For verifiability, we measure mean class accuracy on hidden signals. In RetinaMNIST with PVT-v2, this accuracy is 20.22% on raw data, confirming no hidden signal, but rises to 70.92% (Undercover) and 94.37% (Ours), proving the model has learned the hidden signal. While BloodMNIST shows lower verifiability, sample imbalance correlation helps resolve ambiguity: -0.02 (Undercover) vs. 0.35 (Ours), showing stronger alignment in our case. Overall, our method effectively verifies whether a model was trained on a protected dataset.

Verification for Segmentation Task: Similar to classification, we evaluate the model using three metrics: Harmlessness, Verifiability, and EFAC. Harmlessness and verifiability are assessed via Dice Similarity Coefficients (DSC) for test data and hidden signals. We generate four hidden signal sets and measure the



(a) Trained on Raw Data (b) Trained on Undercover (b) Trained on PRADA

Fig. 4. t-SNE visualization for DermaMNIST case.

correlation between their Dice coefficients and predefined occurrence probabilities. Figure 3 presents examples of segmentation results from a model trained on the PRADA dataset. As shown, the model performs well on raw test data, demonstrating harmlessness. For the least and most frequent hidden signals, the predicted mask is partially broken, but the most frequent signal produces a more defined output, indicating effectiveness of EFAC. As shown in Table 2, the proposed method achieves higher verifiability and EFAC values while maintaining similar test Dice degradation compared to Undercover Bias. This demonstrates that the PRADA dataset remains reliably verifiable even when mixed with other datasets, highlighting its practical applicability for real-world usage.

3.3 t-SNE Visualizations

To analyze how hidden signals operate in trained models, we visualize the feature distribution using t-SNE. Figure 4 shows the t-SNE results for three PVTv2 trained on each dataset (Raw, Undercover, and Ours). We extract features of hidden signals and visualize their distributions, with each color representing a hidden signal class label. As seen in Figure 4(a), the model trained on raw data (without hidden signals) shows no discriminability. However, models trained with Undercover (Figure 4(b)) and Ours (Figure 4(c)) exhibit distinct clusters, indicating that the hidden signals are learned. Moreover, Undercover does not differentiate between frequent and rare cases, treating both equally. In contrast, our method distinctly separates rare and frequent cases, enabling more detailed and robust verification based on signal frequency.

4 Conclusion

In this paper, we propose a two-step dataset verification method that embeds hidden signals while preserving model performance. Building on prior approaches, we leverage data imbalance—traditionally seen as a drawback in training—to improve verification reliability through asymmetric frequency distributions. By intentionally introducing data imbalance and incorporating EFAC, a metric that measures alignment between model performance and predefined occurrence frequency, we enhance dataset misuse detection. Experiments on classification and segmentation tasks demonstrate minimal performance degradation while ensuring strong verifiability through mean class accuracy on hidden signals and EFAC. Our method provides a reliable solution for dataset verification, particularly in medical imaging, where models are highly sensitive to degradation and often constrained by a limited number of classes.

Acknowledgments. This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00334321)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aghakhani, H., Meng, D., Wang, Y.X., Kruegel, C., Vigna, G.: Bullseye polytope: A scalable clean-label poisoning attack with improved transferability. In: EuroS&P. pp. 159–178 (2021)
2. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International conference on machine learning. pp. 233–242. PMLR (2017)
3. Baluja, S.: Hiding images in plain sight: Deep steganography. In: NeurIPS. pp. 2066–2076 (2017)
4. Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., Ma, Y.: Sft memorizes, rl generalizes: A comparative study of foundation model post-training. arXiv preprint arXiv:2501.17161 (2025)
5. Fumero, F., Sigut, J., Alayón, Silvia andGonzález-Hernández, M., González de la Rosa, M.: Interactive tool and database for optic disc and cupsegmentation of stereo and monocular retinal fundus images (06 2015)
6. Geiping, J., Fowl, L.H., Huang, W.R., Czaja, W., Taylor, G., Moeller, M., Goldstein, T.: Witches' brew: Industrial scale data poisoning via gradient matching. In: ICLR (2021)
7. Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S.: Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access **7**, 47230–47244 (2019)
8. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)

9. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
10. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5375–5384 (2016)
11. Huang, W.R., Geiping, J., Fowl, L., Taylor, G., Goldstein, T.: Metapoison: Practical general-purpose clean-label data poisoning. NeurIPS **33**, 12080–12091 (2020)
12. Jang, J., Han, B., Kim, J., Youn, C.H.: Rethinking data bias: Dataset copyright protection via embedding class-wise hidden bias. In: European Conference on Computer Vision. pp. 1–18. Springer (2024)
13. Lan, H., Gu, J., Torr, P., Zhao, H.: Influencer backdoor attack on semantic segmentation. In: The Twelfth International Conference on Learning Representations
14. Li, Y., Bai, Y., Jiang, Y., Yang, Y., Xia, S.T., Li, B.: Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. In: NeurIPS (2022)
15. Li, Z., Kamnitsas, K., Glocker, B.: Analyzing overfitting under class imbalance in neural networks for image segmentation. IEEE transactions on medical imaging **40**(3), 1065–1077 (2020)
16. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2537–2546 (2019)
17. Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis **59**, 101570 (2020)
18. Provos, N., Honeyman, P.: Hide and seek An introduction to steganography. IEEE security & privacy **1**(3), 32–44 (2003)
19. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
20. Saha, A., Subramanya, A., Pirsiavash, H.: Hidden trigger backdoor attacks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11957–11965 (2020)
21. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR (2018)
22. Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J.P., Goldstein, T.: Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In: ICML. pp. 9389–9398 (2021)
23. Shafahi, A., Huang, W.R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., Goldstein, T.: Poison frogs! targeted clean-label poisoning attacks on neural networks. NeurIPS **31** (2018)
24. Sivaswamy, J., Krishnadas, S., Joshi, G.D., Jain, M., Tabish, A.U.S.: Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In: 2014 IEEE 11th international symposium on biomedical imaging (ISBI). pp. 53–56. IEEE (2014)
25. Souri, H., Fowl, L., Chellappa, R., Goldblum, M., Goldstein, T.: Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. NeurIPS **35**, 19165–19178 (2022)
26. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media pp. 1–10 (2022)

27. Wu, D.C., Tsai, W.H.: A steganographic method for images by pixel-value differencing. Pattern recognition letters **24**(9-10), 1613–1626 (2003)
28. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
29. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
30. Zhang, Z., Yin, F., Liu, J., Wong, W., Tan, N., Lee, B., Cheng, J., Wong, T.: Origa: An online retinal fundus image database for glaucoma analysis and research. In: Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. pp. 3065–3068