Assessing Risk of Stealing Proprietary Models for Medical Imaging Tasks

Ankita Raj[1], Harsh Swaika[1], Deepankar Varma[2⋆], and Chetan Arora[1⋆⋆]

[1] Indian Institute of Technology Delhi, India
[2] Thapar Institute of Engineering and Technology, Patiala, India

Abstract. The success of deep learning in medical imaging applications has led several companies to deploy proprietary models in diagnostic workflows, offering monetized services. Even though model weights are hidden to protect the intellectual property of the service provider, these models are exposed to model stealing (MS) attacks, where adversaries can clone the model's functionality by querying it with a proxy dataset and training a thief model on the acquired predictions. While extensively studied on general vision tasks, the susceptibility of medical imaging models to MS attacks remains inadequately explored. This paper investigates the vulnerability of black-box medical imaging models to MS attacks under realistic conditions where the adversary lacks access to the victim model's training data and operates with limited query budgets. We demonstrate that adversaries can effectively execute MS attacks by using publicly available datasets. To further enhance MS capabilities with limited query budgets, we propose a two-step model stealing approach termed `QueryWise`. This method capitalizes on unlabeled data obtained from a proxy distribution to train the thief model without incurring additional queries. Evaluation on two medical imaging models for Gallbladder Cancer and COVID-19 classification substantiate the effectiveness of the proposed attack. The source code is available at https://github.com/rajankita/QueryWise.

Keywords: Model Stealing · Model Extraction · Privacy attack.

1 Introduction

Adversarial attacks. Deep learning models have emerged as a prominent tool in medical imaging, improving diagnosis and accelerating decision-making in clinical tasks. However, these systems pose security and privacy risks. While threats like adversarial [12] and poisoning attacks [20] have been investigated in medical imaging [31,3,24], model stealing attacks remain largely unexplored.

Model stealing attacks. Model Stealing (MS) attacks, also known as Model Extraction attacks, pose a significant threat to the confidentiality of proprietary

⋆ Work done as an intern at Indian Institute of Technology Delhi
⋆⋆ Corresponding author

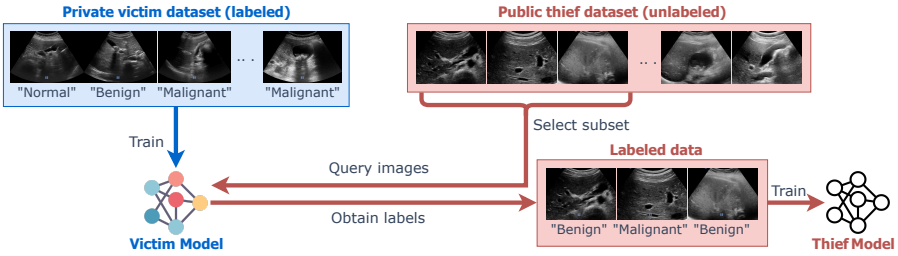

Fig. 1. Model Stealing Attack setup: Victim model is a black-box. An adversary samples images from an unlabeled thief distribution and queries labels from the victim model to train a thief model.

machine learning models [17,21,23]. In an MS attack, adversaries aim to replicate the functionality of a private machine learning model that operates as a black box and is accessible only through querying, while its weights are hidden from end users. This operational model, known as Machine-learning-as-a-Service (MLaaS), has been adopted by several companies [1,2], which provide proprietary machine-learning models to hospitals, healthcare professionals, or end-users for a fee. To safeguard their intellectual property, these companies restrict access to the model's internals, such as the training dataset, model architecture, and hyperparameters, while offering only black-box access to the model. However, in an MS attack, a malicious user tries to replicate the functionality of the black-box model, referred to as the *victim* model, by querying it with a proxy dataset and training a substitute model using the acquired predictions. Given the high labeling costs in medical imaging, often requiring expert knowledge, such attacks could enable competitors to develop their own models at significantly lower costs. This not only jeopardizes the intellectual property of the model owner but also exposes the model to additional threats, such as model inversion attacks.

Challenges with stealing medical imaging models. Despite the significant threat, MS attacks remain under-explored for medical imaging. Existing methods for MS attacks on natural image-based models [21,23,30,7] require thousands or even millions of queries, impractical for medical datasets with limited samples. While some studies have demonstrated attacks on medical datasets, they are either concerned with stealing text-based models [32] or rely on relaxed assumptions necessitating the acquisition of the full prediction vector from the victim model [21]. Our work involves stealing medical imaging models under a more realistic scenario where a victim model provides only the top-1 prediction, also known as *hard-label*, and an extremely small query budget. Under these stringent stealing conditions, the limited queried subset fails to adequately cover the victim's input space, resulting in thief models significantly trailing behind in accuracy compared to the victim.

Our proposal. We note that state-of-the-art (SOTA) MS techniques primarily focus on training the thief model solely on the small queried subset, neglecting

a large portion of unused proxy data. We propose **QueryWise**, a novel MS method that leverages both queried/labeled data and the remaining unlabeled data for thief model training. In the first step, we train a model exclusively on the labeled data, which we call the *anchor model*, using established MS attack techniques [21,23]. The final thief model is trained in the second step, utilizing both labeled and unlabeled data, with guidance from the anchor model and an additional *teacher model*. The teacher model's weights are updated using an exponential moving average of the thief model's weights to ensure stable predictions [28]. Since the thief's proxy data lies out-of-distribution (OOD) with respect to victim model's distribution, hard-labels queried from the victim may lack meaningful information and thus have limited utility. A key insight of our work is that by generating soft pseudo-labels for the unlabeled data using a combination of the anchor and teacher models, our approach facilitates the thief model in capturing label correlations with other samples more effectively. [15] showed the benefits of using unlabeled data for MS attacks, but it remained unclear if these benefits applied when the proxy data differed from the victim data. [30] employed an iterative training method with random erasure to generate pseudo-labels for unlabeled data. Our method improves upon this by combining pseudo-labels from an anchor model and a weighted-averaged teacher model, retaining only high-confidence labels for superior supervision. Crucially, we use the anchor model to guide the thief model on labeled data (see Section 3.2).

Contributions. We make the following key contributions: **(1)** We formally study MS attacks for medical image classification models under a realistic threat model of 5000 queries and hard-label access. **(2)** We propose a novel MS method for training a thief model leveraging both labeled and unlabeled data. We effectively utilize supervision from an anchor model trained exclusively on labeled data, and a teacher model to make use of the thief model's unlabeled data. This ensures our technique successfully mounts MS attacks with low query budgets and without access to data from the victim model's distribution. **(3)** Our evaluation of SOTA model stealing defenses reveals their failure to consistently defend against different types of MS attacks, underscoring the pressing need to address the serious implications of MS attacks in medical imaging.

2 Threat Model

Hard labels. We investigate attacks on victim models f_{V} trained on data distribution $P_V(X)$ for medical image classification. The output of the model $y \in \{1, \dots K\}$ is a distribution over K classes. The attacker has black-box query access to f_{V}, allowing it to submit an image x and receive the models's prediction. Unlike many MS attacks, which assume access to the full probability vector $f_V(x) \in \mathbb{R}^K$, we consider a stricter scenario where only the topmost prediction $\mathrm{argmax}_{i \in \{1,\dots,K\}} f_V(x)_i$ from the victim is available. Additionally, the attacker lacks knowledge of the victim model's hyperparameters.

Different thief and victim data distributions. Most MLaaS medical imaging models are trained on confidential patient data inaccessible to adversaries.

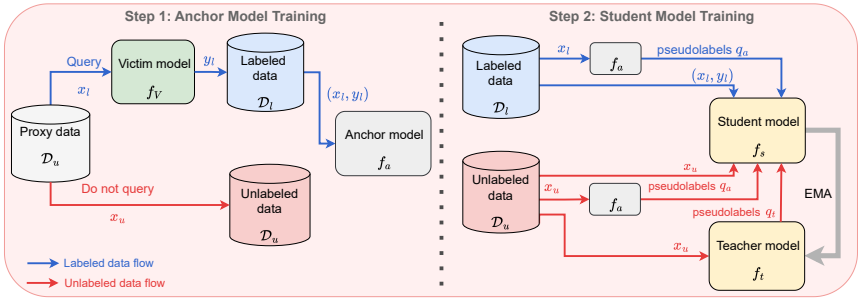

Fig. 2. Overview of QueryWise: First, a subset of the proxy data is queried from the victim model, and an *anchor* model is trained on the labeled data. Our main contribution lies in the second step, where a *student* model is trained using both labeled and unlabeled data. On labeled data, the student receives supervision from hard labels obtained from the victim and soft labels generated by the anchor model. On unlabeled data, pseudolabels are generated using both the anchor model and a *teacher* model (which is updated at the end of each iteration from the student model using exponential moving average). The student model serves as the final thief model.

Consequently, thief models must be trained on proxy datasets. While typical MS attacks [21,23] use large-scale public datasets of natural images like ImageNet [25], these datasets are ill-suited for our task due to their dissimilarity to medical images. Instead, malicious actors can exploit freely available medical imaging datasets online for stealing models. Therefore, a more natural choice for attackers is to use a dataset $P_A(X)$ of publicly accessible unlabeled images from the same modality. The attacker selects a subset of images from $P_A(X)$ which it queries from the victim model, thus constructing a labeled set $\mathcal{D}_l = \{(x_l^{(i)}, y_l^{(i)})\}_{i=1}^{N_l}$ of size N_l which is then used to train the thief model f_T.

3 Proposed Method

Figure 2 shows an overview of the proposed model stealing method. In addition to the conventional MS process of querying the proxy data and training the thief model on the labeled set, we additionally train on the unlabeled data as well. The proposed stealing process is carried out in two steps, as discussed below.

3.1 Query set selection and Anchor model training

In the first step, we train a fully supervised thief model using the limited labeled data $\mathcal{D}_l$, and call it the anchor model, f_a. Note that the thief only needs to curate unlabeled images; these are labeled later by querying the victim model. Existing MS attacks use concepts like active learning [23] and reinforcement learning [21] to iteratively select the query set and train the thief model. We re-purpose existing techniques to train our anchor model.

3.2 Student model training

The main contribution of our method lies in the second step, which is the training of student model. Our strategy is inspired by Knowledge Distillation [14], but extends it to incorporate training using labeled as well as unlabeled data. Knowledge from labeled data is embedded in the fixed anchor model trained in the first step, while an additional teacher model, described below, incorporates knowledge from labeled and unlabeled data both, and is dynamically updated with the student. This additional training on the unlabeled data using artificially generated pseudo-labels (by anchor and teacher models) helps the student model surpass the anchor's performance, as can be verified from Table 1 in our experiments. Given a training mini-batch comprising B_l labeled samples and B_u unlabeled samples, an aggregate loss $\mathcal{L} = \mathcal{L}_l + \lambda \mathcal{L}_u$ is computed by combining labeled and unlabeled losses, as discussed below.

Labeled loss, $\mathcal{L}_l$. The loss is computed from the labeled samples in a mini-batch. It is a weighted combination of cross-entropy loss $\mathcal{L}_{\text{CE}}$ computed from the hard labels queried from the victim, and knowledge distillation loss $\mathcal{L}_{\text{KD}}$ computed from the anchor model's output probability vectors on the labeled data. This extra supervision from the anchor model provides regularization benefits similar to those provided by Self-Distillation [11].

$$\mathcal{L}_l = (1-\alpha)\mathcal{L}_{\text{CE}} + \alpha \mathcal{L}_{\text{KD}}, \qquad \text{where} \tag{1}$$

$$\mathcal{L}_{\text{CE}} = \sum_{i=1}^{B_l} \mathcal{H}\big(\sigma(q_s^i), y_l^i\big), \qquad \text{and} \tag{2}$$

$$\mathcal{L}_{\text{KD}} = \sum_{i=1}^{B_l} \tau^2 \ \text{KL}\left(\sigma\left(\frac{q_s^i}{\tau}\right), \sigma\left(\frac{q_a^i}{\tau}\right)\right). \tag{3}$$

Here y_l^i are the hard labels queried from the victim model, $q_s^i = f_s(x_l^i)$ and $q_a^i = f_a(x_l^i)$ denote the logits produced by the student and anchor models respectively for the labeled data, and $\sigma(\cdot)$ is the softmax function. $\mathcal{H}$ is the cross-entropy loss and $\text{KL}(\cdot,\cdot)$ is KL-divergence loss, with distillation temperature [14] $\tau > 1$ used to smoothen the anchor's predictions. $\alpha \in (0, 1]$ controls the relative importance of the two losses.

Unlabeled loss, $\mathcal{L}_u$. The unlabeled loss tries to match the student's predictions with the softened softmax outputs of both the teacher model and the anchor model via a weighted sum of two knowledge-distillation loss components: $\mathcal{L}_{\text{KD}}^t$ computed using the teacher model, and $\mathcal{L}_{\text{KD}}^a$ using the anchor model.

$$\mathcal{L}_u = (1-\beta)\mathcal{L}_{\text{KD}}^t + \beta \mathcal{L}_{\text{KD}}^a, \qquad \text{where} \tag{4}$$

$$\mathcal{L}_{\text{KD}}^t = \sum_{i=1}^{B_u} \mathbb{1}\left(\sigma(q_a^i) > \rho\right) \tau^2 \ \text{KL}\left(\sigma\left(\frac{q_s^i}{\tau}\right), \sigma\left(\frac{q_t^i}{\tau}\right)\right), \qquad \text{and} \tag{5}$$

$$\mathcal{L}_{\text{KD}}^a = \sum_{i=1}^{B_u} \mathbb{1}\left(\sigma(q_a^i) > \rho\right) \tau^2 \ \text{KL}\left(\sigma\left(\frac{q_s^i}{\tau}\right), \sigma\left(\frac{q_a^i}{\tau}\right)\right). \tag{6}$$

Here $q_s^i = f_s(x_u^i)$, $q_t^i = f_t(x_u^i)$, $q_a^i = f_a(x_u^i)$ denote the logits produced by the student, teacher and anchor models respectively for the unlabeled data. The relative contribution of the two KD loss components from the teacher and anchor is controlled by the hyper-parameter $\beta \in (0, 1]$. To curb the effect of noisy pseudo-labels [26], only samples with the maximum softmax probability above a threshold ρ are allowed to contribute to the unlabeled loss.

Role of teacher model. The teacher model is initialized from the weights of the anchor model, and after each mini-batch, it is updated using an exponential moving average of the student's weights [28], defined by $\theta_t \leftarrow m\theta_t + (1 - m)\theta_s$ where θ_s and θ_t are the student's and teacher's weights respectively, and m is the smoothing factor. The motivation for introducing the teacher model is to prevent the student from confirmation bias on a particular batch. At the same time, the teacher model also escapes rigidity of the anchor model by slowly updating itself from the student model based on the training on unlabeled data. The anchor and teacher models thus complement each other, and represent a balance between exploitation and cautious exploration in our method. Recall that at the end of the training procedure, thief model is derived from the student model.

Handling class imbalance. We observe that the labeled set obtained by querying the victim model on an out-of-distribution thief dataset is often class-imbalanced. This may result in the thief model generalizing poorly on the less-frequent classes. To encourage equal contribution from all classes, we incorporate Logit Adjustment (LA) [19] by applying label-dependent offsets to the student's logits q_s^i during computation of the labeled loss $\mathcal{L}_l$ while training the student.

4 Experiments and Results

Model stealing setup. We use two victim models to demonstrate our model stealing capabilities. **(1)** The first is **RadFormer** [5], a transformer-based model for classification of **Gallbladder Cancer** (GBC) from Ultrasound (US) images. It is trained on the Gallbladder Cancer Ultrasound (GBSU) dataset [4] comprising of 1255 US images from 218 patients, categorized into 432 normal, 558 benign, and 265 malignant images. We use the publicly available model shared by the authors as our victim model. For stealing, we use the publicly available GBC US videos dataset [6], consisting of 32 malignant and 32 non-malignant videos containing 12,251 and 3,549 frames respectively, as our proxy dataset. Note that we do not use the video labels, but instead query the frames from the victim model and use the acquired labels instead. **(2)** The second victim model is a ResNet18 trained on **POCUS** dataset [8] for **COVID-19 classification** that consists of 2116 images, of which 655, 349, and and 1112 are of COVID-19, bacterial pneumonia, and healthy control respectively. A subset of the public COVIDx-US dataset [10] comprising 13032 lung US images is used as the proxy dataset. For both tasks, we use a query budget of 5000 samples. Note that a budget of 5000 is much lower in comparison to existing MS attacks on general vision tasks [23,7] that require millions of queries. Further, model stealing requires only

Table 1. Model stealing performance for GBC malignancy classification task. We report accuracy (Acc.), specificity, sensitivity and agreement (Agr.). Query budget is 5000. In a significant achievement, we note that the thief model is able to outperform the radiologist accuracy using the proposed MS attack.

Arch	Method	Acc.	Spec.	Sens.	Agr.
Custom [4]	Victim	90.16	90.00	92.86	-
	Radiologist A [4]	70.00	87.30	70.70	-
	Radiologist B [4]	68.30	81.10	73.20	-
ResNet50 [13]	Random [21]	66.39	85.00	61.90	71.31
	k-Center [23]	71.31	87.50	71.43	68.85
	Random+FixMatch [26]	65.57	82.00	62.00	66.39
	Random+QW	71.31	80.00	**81.00**	74.59
	k-Center+QW	**72.95**	79.00	**81.00**	80.33
Inception-v3 [27]	Random [21]	69.67	71.25	**83.33**	71.31
	Random+QW	**70.49**	74.00	74.00	72.13
ViT [9]	Random [21]	62.30	81.25	66.67	67.21
	Random+QW	**65.57**	76.00	**69.00**	72.13
DeiT [29]	Random [21]	71.31	81.25	78.57	74.59
	Random+QW	**77.05**	76.00	**90.00**	77.05

unlabeled images for querying, which are easier to curate than labeled images, even in medical settings.

Training details. We train our thief models using a query budget of 5000 samples, of which 10% samples are set aside as validation data for hyper-parameter selection, and the rest are used for training. QueryWise training incorporates two stages. **First stage:** The anchor model in the first stage can be obtained from any of the baseline MS methods. In our experiments, we obtain the anchor model using two baseline methods: KnockoffNets' Random selection [21] and ActiveThief's k-Center method [23]. k-Center training is done for 5 cycles, with the training budget split uniformly across cycles. Following [23], we evaluate the F1 score on the validation set at the end of each epoch, and use these scores to select the best model for each k-Center training cycle. **Second stage:** The student model in the second stage is initialized from ImageNet-pretrained weights and trained using the same labeled set (of size 5000, queried from the victim in first stage), and train-val split as used by the anchor. We train the student model for 100 epochs (where one epoch is defined as one pass over the labeled data, while the remaining unlabeled data is spilled over to the next epoch). We use $\alpha = 0.4$, $\beta = 0.5$, $T = 1.5$, $\lambda = 1$, $\rho = 0.95$, and $m = 0.999$. Other training hyperparameters are listed in Supplementary Table S4.

Comparison with Baselines. We compare QueryWise (QW) to existing MS techniques: Knockoff Nets' Random selection [21], and k-Center from ActiveThief [23]. We report model stealing performance for the GBC malignancy classification task in Table 1. In addition to the standard evaluation metrics of accuracy, specificity and sensitivity, we report *agreement*, which measures how often the thief's prediction matches the victim's. For ResNet50 thief architecture, we im-

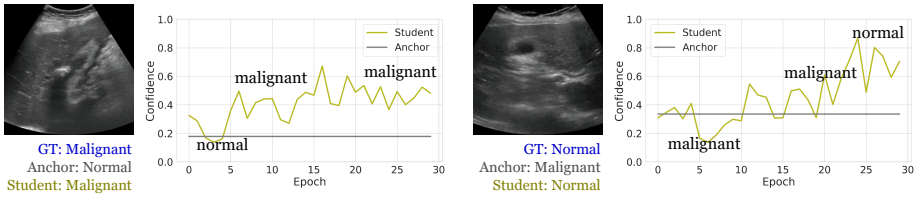

Fig. 3. Unlabeled data improves thief accuracy: Comparing predictions of an *anchor* model with a *student* model on two GBC test set images, we show the evolution of student model's confidence for the ground truth (GT) class over the course of training, along with predicted labels. Anchor model's confidence is also shown for reference. Initially aligned with the anchor, the student gains confidence in the GT class with the use of unlabeled data, eventually making correct predictions.

plement QueryWise alongside two anchor models obtained from Random and k-Center MS attacks respectively. It can be seen that proposed method outperforms the respective anchor models, as well as a popular semi-supervised learning technique FixMatch [26], in terms of both accuracy and sensitivity. We also evaluate the impact of varying the thief architecture, and observe that transformer-based networks like DeiT [29] increases model stealing accuracy as well as sensitivity. Given a fixed architecture, the proposed method (QW+Random) outperforms the respective anchor model (Random) in terms of both accuracy and agreement. **In a siginificant achievement, we show that a thief model using the proposed model stealing attack is able to outperform Radiologists' accuracy (70% against 77% using our MS attack), indicating the serious threat posed by the proposed attack on the proprietary models.** We report additional results on the model stealing performance for the COVID-19 classification task in supplementary Table S1, and for natural image classifiers (trained on datasets like CIFAR-10 and Caltech-256) in supplementary Table S2, demonstrating the generality of our method across victim models.

Qualitative Comparison. Figure 3 shows the role of unlabeled data in improving thief model accuracy. We compare the predictions of anchor model with the student model, and show instances where the latter overcomes the mistakes made by the former model by learning from unlabeled data.

Effectiveness of Defenses. We evaluate MS attacks on medical imaging models against SOTA MS defense techniques, in particular perturbation-based defenses that modify the victim model's output vector to prevent an attacker from accurately copying the model. These defenses trade-off victim model's accuracy for security by perturbing the predictions of the model. Three defense techniques are evaluated: Maximum Angular Deviation (MAD) [22] ($\epsilon = 0.8$), Adaptive Misinformation (AM) [16] with detection threshold $\tau = 0.7$ and Gradient Redirection (GRAD^2) [18] with $\epsilon = 0.8$. We observe that none of the evaluated defenses consistently reduce thief accuracy for all MS methods. Moreover, any drop in thief accuracy is usually accompanied by a drop in victim accuracy, thus questioning the utility of these defenses. Detailed results are in suppl. Table S3.

5 Conclusion

We investigated the susceptibility of deployed deep-learning medical imaging models to model stealing attacks within realistic constraints, including limited query budgets and hard-label access. We proposed a new query-efficient method that effectively utilizes unlabeled data from the thief's proxy dataset alongside labeled data queries to enhance thief model performance under low query budgets. Our method proves superior to existing model stealing baselines at stealing medical imaging models. Our research invalidates the common belief regarding the safety of MLaaS framework to prevent against theft of proprietary information, and highlights the need for robust defenses against model stealing attacks.

Acknowledgments. We acknowledge and thank the funding support from AIIMS Delhi-IIT Delhi Center of Excellence in AI funded by Ministry of Education, government of India, Central Project Management Unit, IIT Jammu with sanction number IITJMU/CPMU-AI/2024/0002. We would also like to thank Prof Vikram Goyal, Akshit Jindal, and Rakshita Choudhary for their valuable inputs to the research.

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

References

1. Qure.ai. https://www.qure.ai/, accessed: 2024-03-02
2. Skinvision. https://www.skinvision.com/, accessed: 2024-03-02
3. Alkhunaizi, N., Kamzolov, D., Takáč, M., Nandakumar, K.: Suppressing poisoning attacks on federated learning for medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 673–683. Springer (2022)
4. Basu, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Surpassing the human accuracy: detecting gallbladder cancer from usg images with curriculum learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20886–20896 (2022)
5. Basu, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Radformer: Transformers with global–local attention for interpretable and accurate gallbladder cancer detection. Medical Image Analysis **83**, 102676 (2023)
6. Basu, S., Singla, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 423–433. Springer (2022)
7. Beetham, J., Kardan, N., Mian, A., Shah, M.: Dual student networks for data-free model stealing. In: Dual Student Networks for Data-Free Model Stealing. Eleventh International Conference on Learning Representations (ICLR) (2023)
8. Born, J., Wiedemann, N., Cossio, M., Buhre, C., Brändle, G., Leidermann, K., Goulet, J., Aujayeb, A., Moor, M., Rieck, B., et al.: Accelerating detection of lung pathologies with explainable ultrasound image analysis. Applied Sciences **11**(2), 672 (2021)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
10. Ebadi, A., Xi, P., MacLean, A., Tremblay, S., Kohli, S., Wong, A.: COVIDx-US - An Open-Access Benchmark Dataset of Ultrasound Imaging Data for AI-Driven COVID-19 Analytics. arXiv:2103.10003 (2021)
11. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: International Conference on Machine Learning. pp. 1607–1616. PMLR (2018)
12. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015)
15. Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., Papernot, N.: High accuracy and high fidelity extraction of neural networks. In: 29th USENIX security symposium (USENIX Security 20). pp. 1345–1362 (2020)
16. Kariyappa, S., Qureshi, M.K.: Defending against model stealing attacks with adaptive misinformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2020)

17. Kumar, R.S.S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., Xia, S.: Adversarial machine learning-industry perspectives. In: 2020 IEEE Security and Privacy Workshops (SPW). pp. 69–75. IEEE (2020)
18. Mazeika, M., Li, B., Forsyth, D.: How to steer your adversary: Targeted and efficient model stealing defenses with gradient redirection. In: International Conference on Machine Learning. pp. 15241–15254. PMLR (2022)
19. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: International Conference on Learning Representations (2020)
20. Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E.C., Roli, F.: Towards poisoning of deep learning algorithms with back-gradient optimization. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. pp. 27–38 (2017)
21. Orekondy, T., Schiele, B., Fritz, M.: Knockoff nets: Stealing functionality of black-box models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4954–4963 (2019)
22. Orekondy, T., Schiele, B., Fritz, M.: Prediction poisoning: Towards defenses against dnn model stealing attacks. In: International Conference on Learning Representations (2019)
23. Pal, S., Gupta, Y., Shukla, A., Kanade, A., Shevade, S., Ganapathy, V.: Activethief: Model extraction using active learning and unannotated public data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 865–872 (2020)
24. Pandey, P., Vardhan, A., Chasmai, M., Sur, T., Lall, B.: Adversarially robust prototypical few-shot segmentation with neural-odes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 77–87. Springer (2022)
25. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
26. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
28. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
29. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
30. Wang, Y., Li, J., Liu, H., Wang, Y., Wu, Y., Huang, F., Ji, R.: Black-box dissector: Towards erasing-based hard-label model stealing attack. In: European Conference on Computer Vision. pp. 192–208. Springer (2022)
31. Xu, M., Zhang, T., Li, Z., Liu, M., Zhang, D.: Towards evaluating the robustness of deep diagnostic models by adversarial attack. Medical Image Analysis **69**, 101977 (2021)
32. Zhang, L., Lin, G., Gao, B., Qin, Z., Tai, Y., Zhang, J.: Neural model stealing attack to smart mobile device on intelligent medical platform. Wireless Communications and Mobile Computing **2020**, 1–10 (2020)

Supplementary Material

Table S1. Model stealing performance for COVID-19 classification task. We report total accuracy (Total), class-wise accuracies for all 3 classes, and agreement (Agr.). Query budget is 5000. Proposed method achieves thief accuracy close to the baselines, while having the best agreement value.

Arch	Method	Total	COVID-19	Pneumonia	Regular	Agr.
Victim	-	89.91	83.43	95.40	92.24	-
ResNet-50[13]	Random [21]	65.97	40.13	74.71	80.17	70.59
	k-Center [23]	65.55	57.32	68.97	69.83	68.49
	Random+QW	63.87	33.12	72.41	81.47	71.22

Table S2. Model extraction performance on general vision tasks for natural images with 5000 queries. The proposed method is implemented with two different anchor models: Random+QW and k-Center+QW. The best method for each dataset is depicted in **bold**, and the next best is underlined. Proposed method outperforms the baselines in terms of both accuracy and agreement for all datasets. Note: Dual Students[7] is a data-free method that uses synthetically generated data instead of a proxy dataset, but requires millions of queries. We implement [7] with a budget of 500K queries for the smaller datasets, yet it fails to match the performance of the other methods operating at 5000 queries.

Method	Venue	MNIST		SVHN		CIFAR10		Caltech256		CUBS200		Indoor67	
		Acc	Agr	Acc	Agr	Acc	Agr	Acc	Agr	Acc	Agr	Acc	Agr
Random[21]	CVPR'19	80.55	80.59	67.54	67.84	65.89	66.66	39.77	40.32	14.74	15.75	33.36	36.22
Entropy[23]	AAAI'20	80.55	80.59	41.02	41.11	47.53	48.16	38.88	39.78	13.87	15.06	35.82	39.33
k-Center[23]	AAAI'20	71.92	71.99	72.78	73.25	68.02	68.71	44.66	45.16	18.57	20.14	40.37	42.91
BBD + Random [30]	ECCV'22	27.49	27.51	58.87	59.04	47.04	47.59	40.56	41.16	16.00	16.64	34.40	38.06
BBD + k-Center [30]	ECCV'22	58.53	58.52	43.99	44.18	40.59	40.58	41.72	42.05	16.48	17.41	30.07	33.88
Dual Students [7]	ICLR'23	18.62	19.24	6.69	10.89	12.86	10.16	-	-	-	-	-	-
Random + QW		80.00	80.08	70.83	71.20	73.07	73.62	45.19	45.36	14.67	15.43	36.12	38.63
k-Center + QW		**85.08**	**86.01**	**76.58**	**76.92**	**74.85**	**74.78**	**50.48**	**50.00**	**20.21**	**21.32**	**42.24**	**43.73**

Table S3. Thief model accuracy under SOTA model stealing defenses. We evaluate three MS attacks on GBC malignancy classification victim model, under three defense techniques. Note that the RadFormer victim model is non-differentiable, rendering it infeasible for the defenses to compute gradients. Hence, for this experiment, we use a differentiable version of RadFormer, containing only the global branch. As can be observed, there is no significant impact (lowering of thief accuracy) that is consistent across all MS attacks. The paper advocates more research in this topic to prevent stealing of proprietary information through this route of MS attacks.

Method	No Defense	MAD [22]	AM [16]	GRAD^2 [18]
Victim model	89.34	80.32	88.52	86.88
Random	75.40	78.68	62.29	69.67
Entropy [23]	74.59	64.75	65.57	75.40
k-Center [23]	70.49	72.13	72.95	79.50

Table S4. Training hyperparameters for anchor and student models, corresponding to the two victim models. B_l and B_u are mini-batch sizes for labeled and unlabeled data respectively. Input image pre-processing for ViT, DeiT and Inception-v3 includes random horizontal flip and random augmentation; for ResNet-50 includes random crop, jitter, and random horizontal flip. For student model training, we use cosine learning rate decay with warmup.

		GBC				COVID-19
		ResNet50	Inception-v3	ViT	DeiT	ResNet50
Anchor training	learning rate	0.005	0.005	0.005	0.005	0.01
	momentum	0.9	0.9	0.9	0.9	0.9
	epochs	100	100	100	100	100
	batch size	16	16	16	16	128
	weight decay	0.0005	0.0005	0.0005	0.0005	0.0005
Student training	learning rate	0.02	0.01	0.01	0.01	0.02
	momentum	0.9	0.9	0.9	0.9	0.9
	weight decay	0.0005	0.0005	0.0005	0.005	0.005
	epochs	100	100	100	100	100
	warmup epochs	10	10	10	10	10
	B_l	16	16	16	16	16
	B_u	112	112	48	48	112