Protecting Deep Cerebrospinal Fluid Cell Imaging Models with Backdoor and Semi-Distillation

Fang-Qi Li[1], Shi-Lin Wang[1,2*], Zhen-Hai Wang[2,3*]
[1] Shanghai Jiao Tong University, School of Electronic Information and Electrical Engineering, Shanghai, China
[2] Diagnosis and Treatment Engineering Technology Research Center of Nervous System Diseases of Ningxia Hui Autonomous Region, Yinchuan, China
[3] Neurology Center, Ningxia Medical University General Hospital, Yinchuan, China
[*]wsl@sjtu.edu.cn, wangzhenhai1938@163.com

***Abstract*—Cerebrospinal fluid imaging models turn out to be a promising computer aided diagnosis technique. Current models can efficiently and correctly identify numerous categories of cells within a slice image of cerebrospinal fluid. Training a cerebrospinal fluid imaging model, especially a deep neural network, requires vast amount of data. Collecting necessary data for medical tasks is an expensive process, during with many experts, devices, and privacy concerns have to be involved. Therefore, it is crucial to protect such models from piracy and reselling. In this paper, we study the problem of intellectual property protection of deep cerebrospinal fluid imaging models. We adopt backdoor-based watermarking as the ownership evidence and propose a semi-distillation framework to embed the watermark into the model. The proposed scheme can verify the ownership of the genuine author, hence provide robust and unforgeable protection over deep cerebrospinal fluid imaging models.**

Keywords—computer aided diagnosis, model protection, cerebrospinal fluid cell imaging, AI security

I. INTRODUCTION

With the development of artificial intelligence (AI) in the computer vision (CV) domain, many AI models have been proposed for medical discipline, where image is the major source of information. Processing medical image had been a tedious job, yet only well-trained medic staff are capable of conducting it. For example, doctors have to observe the patient's lung from the chest radiographs. In leukemia diagnosis, doctors have to count the number of leukocyte from the patient's spinal fluid. The task of analyzing cerebrospinal fluid images is an important emerging challenge. On one hand, cerebrospinal fluid turns out to be an informative source of diagnosis, from with some difficult illnesses were finally identified [1]. On the other hand, the number of categories of cells appear in cerebrospinal fluid is large, and the density of cells varies across different domains. These facts increase the difficulty in designing AI models for cerebrospinal fluid image processing, especially cell segmentation and counting.

Unlike traditional CV discipline in which the performance of models is evaluated on established datasets, in computer aided diagnosis (CAD), the datasets are hardly published [2][3][4]. The reasons behind this phenomenon are:

- Collecting and labelling medic images is much more expensive than general images of animals, faces, or appliances. Since the former has to be assisted by medical equipment and experts. Therefore, few facilities are willing to share their data for free.

- The medical images are inherently involved with patient's privacy. Unauthorized publishing might turn out to be a threat to patient's interest.

Although images are hidden, the CAD models can be published and shared among facilities to boost the medical service [2][3][4]. However, a CAD model, e.g., a deep cerebrospinal fluid imaging model, is confronted by various types of model piracy. Since CAD models contain the contributions of many parties involved in data collecting, processing, and model tuning, they should be protected as intellectual properties. Protecting AI models is usually realized by watermarking, during with the evidence of the author's identity is embedded into the model. Although there have been numerous watermarking schemes for ordinary image classification models, it remains difficult to directly apply existing watermarking schemes onto deep cerebrospinal fluid imaging models for three reasons:

- Medical models are uniformly deployed as API services, so a majority of the current watermarking schemes, namely the white-box watermarking schemes, cannot be adopted.

- The output of deep cerebrospinal fluid imaging models consists of cell segmentation, classification and counting, among which traditional backdoor cannot be inserted, so most black-box watermarking schemes fail as well.

- Deep cerebrospinal fluid imaging models have heavy post-processing modules, so many watermarking schemes that significantly modify the backbone structure might damage the models' overall performance.

Given those challenges, we propose a unified framework for protecting deep cerebrospinal fluid imaging model by watermarking. To enable piracy detection in the worst case (i.e. the author/notary only has the black-box access to the pirated model), we insert backdoor as the watermark. To minimize the influence of the watermark on the model's performance, we design a semi-distillation paradigm, during which a part of the

backbone neural network architecture is tuned to learn the backdoor without forgetting the normal images. The contribution of this paper is three-folded:

• We formulate the threat model for intellectual property protection of CAD models.

• We design a watermarking scheme for deep cerebrospinal fluid imaging model.

• Experiments demonstrated the correctness and reliability of our proposal.

The paper proceeds as follows: in Section II we introduce the backgrounds of deep cerebrospinal fluid cell imaging model and deep neural network watermarking schemes. In Section III we propose our method. Experiments and discussions are provided in Section IV. Finally, Section V concludes the paper.

II. BACKGROUNDS AND RELATED WORKS

A. Deep cerebrospinal fluid cell imaging model

As shown in Fig. 1, a typical cerebrospinal fluid cell image contains numerous categories of cells including: erythrocyte, leukocyte, lymphocyte, etc.

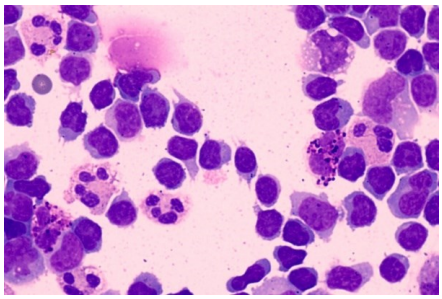

Figure 1. A cerebrospinal fluid cell image.

The number of these cells is a representative indicator of certain diseases. Since identifying and counting cells from cerebrospinal fluid slice is expensive regarding manual effort, many AI models, especially deep neural network (DNN) based models, have been designed to automatize this procedure. A deep cerebrospinal fluid imaging model usually consists of three modules: the convolutional neural network (CNN) backbone, the predictor module, and the post-processing module [5].

Given a cerebrospinal fluid image, the CNN backbone firstly maps it into numerical features. Then the predictor module locates and identifies the cells, and besieges candidate bounding boxes around them. Finally, the post-processing module selects the optimal bounding boxes, counts the number of each category of cells and returns the report. The entire process is demonstrated in Fig. 2.

The CNN backbone is usually instantiated as the classical neural network architecture in CV tasks, e.g., residual networks as ResNet-50, ResNet-101 [6]. As for the predictor modules, Fast RCNN is usually adopted as the bounding box predictor while Cascade Mask RCNN is usually adopted as the mask predictor [7]. The bounding box predictor locates bounding boxes around targets, while the mask predictor is in charge of identifying the pixels that belong to the object of interest. These components are uniformly borrowed directly from the CV community, since object detection and semantic segmentation of images have been studied for a long time. Meanwhile, the post-processing module for cerebrospinal fluid cell image analysis requires extra effort to design. Unlike ordinary images, cerebrospinal fluid cell images are usually extensively crowded, i.e., many cells appear densely within a small region inside the image. Even though there are methods specifically designed for crowded images, they can hardly adapt to cerebrospinal fluid cell images. Traditional post-processing methods such as the Non-Maximum-Suppression (NMS) might decrease the model's overall performance by miscalculating the number of cells. To cope with this dilemma, Hierarchy-NMS (H-NMS) [5] has been proposed to merge candidate bounding boxes in a more intelligent manner. By making use of the semantics within the hierarchy structure in the candidate bounding boxes for cerebrospinal fluid cell images, H-NMS significantly increases the backend performance of current deep cerebrospinal fluid cell imaging models.

B. Deep neural network watermarking

To verify the author of a DNN model, researchers adopted watermark, which is designed to protect the multimedia objects as intellectual properties. The reasons behind this choice are (1)

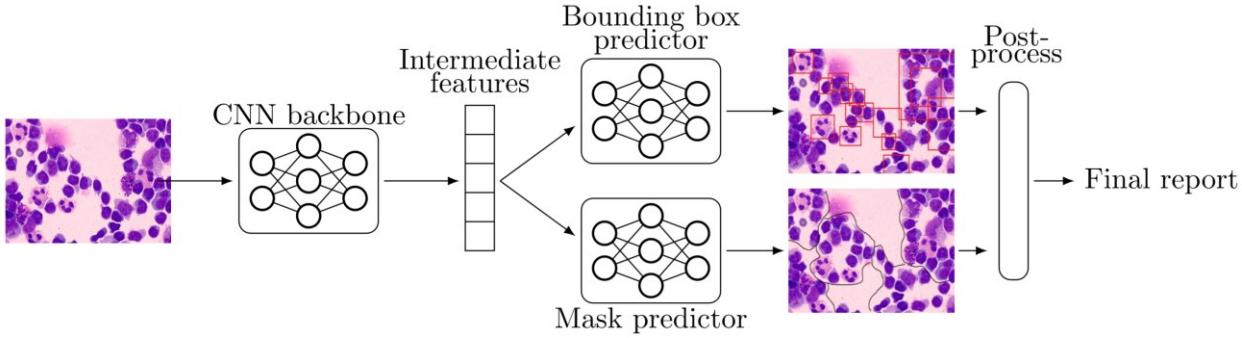

Figure 2. A deep cerebrospinal fluid cell imaging model.

like multimedia objects, DNN models are usually uploaded onto public channels. (2) DNN models are semantically invariant under slight distortion as other multimedia objects.

Watermarking schemes for DNN model can be classified into two categories: the white-box schemes and the black-box schemes. If the author and the notary have white-box access to the possibly stolen model, then the watermark can be encoded into the model's weights and intermediate outputs [8], such schemes are known as the white-box schemes. Otherwise, when the author and the notary can only interact with the suspicious model with an oracle/API then the backdoor-based watermarking schemes are preferred. Backdoor is originally an attack against DNN, where some triggers belong to the input domain evoke specific outputs that are beyond the original task. Conversely, the author can insert specific backdoor into its DNN model as its identity proof [9][10].

III. The Proposed Method

A. Motivation

We consider the black-box setting as the underlying threat model. On one hand, DNN models for medical are usually deployed as online API services. On the other hand, the black-box threat model is strictly stronger than the white-box counterpart, hence the model's security under this scenario is more challenging. We design a backdoor-based watermarking scheme for deep cerebrospinal fluid imaging models. Since the post-processing modules of such models are usually fixed algorithms from which no gradient can be returned, we insert the backdoor into the intermediate output of the model. In this way, the triggers first evoke specific outputs from the CNN backbone, and thence the final outputs. In order to reduce the impact of the watermark on the model's overall performance, we propose a two-stage training process. The first stage is the normal training phase during which the entire model learns the correct labeled and annotated images. While in the second phase, a semi-distillation procedure is adopted to embed the triggers into the intermediate configurable layers while preserving the DNN's normal behavior.

B. Details

The overall framework is illustrated in Fig. 3.

Recall that the network architecture consists of a CNN backbone, a bounding box predictor, a mask predictor, and a

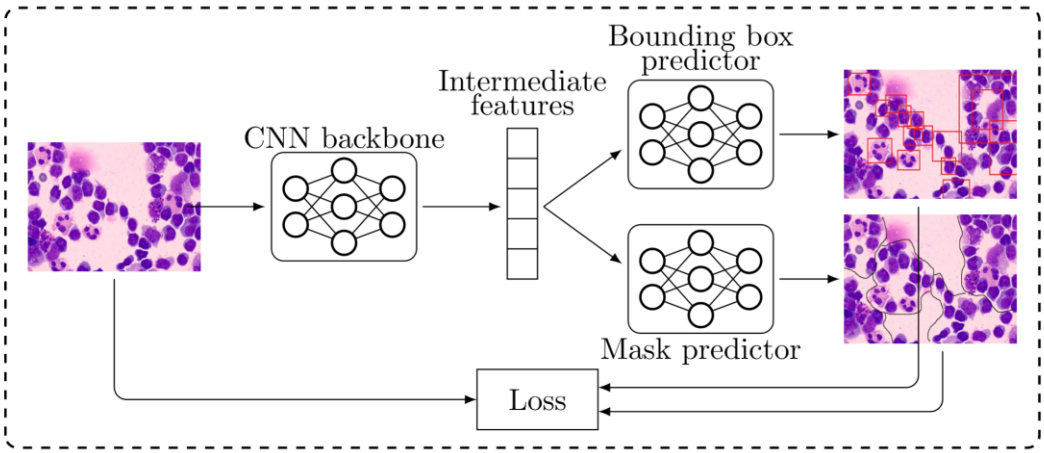

Figure 3 (a). The normal training process.

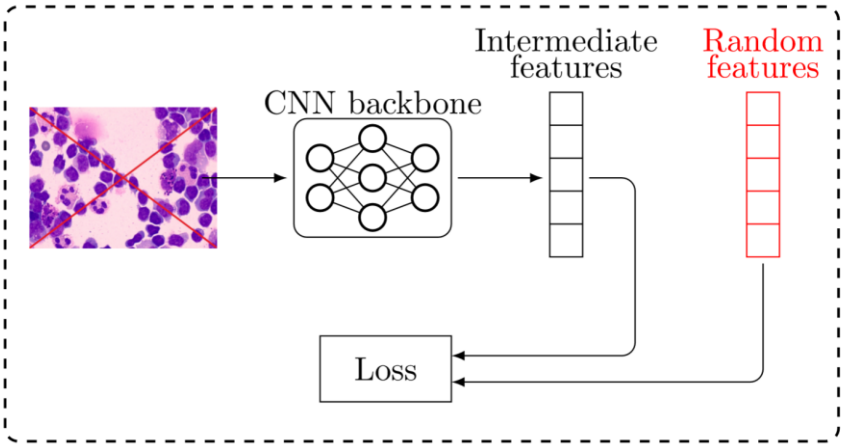

Figure 3 (b). The watermark embedding process by semi-distillation.

post-processing module. In the first phase illustrated by Fig. 3 (a), the network is trained to minimize the joint loss with: bounding box location loss, mask location loss, classification loss, etc.

$$L_{\text{normal}}(\text{W}) = \sum_{x,y} L_{BB}(x, y|W) + L_M(x, y|W),$$

where *(x,y)* denotes an image with its label, *W* is the parameters within the entire model, L_{BB} and L_M are the loss from bounding box location and mask location respectively. The loss is evaluated as a function of *W*. Once such computing graph is established, back-propagation can be conducted to reduce the loss w.r.t. *W*. This gradient-descend process results in a clean DNN model for cerebrospinal fluid image analysis.

In the second phase illustrated by Fig. 3 (b), we freeze the parameters in the predictors and feed trigger images into the CNN backbone, it is expected that the outputs of the CNN backbone on triggers are randomized/null, while those on normal images remain the same as in the first phase. Therefore, the second phase is tantamount to minimize the following loss:

$$L_{\text{watermark}}(W) = L_{\text{normal}}(W) + \lambda \cdot \sum_{x',z'} L_2\big(z', \text{CNN}(x'|W)\big),$$

where λ is the regularizing factor, *(x'z')* denotes a trigger and a randomly generated intermediate feature. The first term of this loss function regulates the deviation between the watermarked model and the clean one. This loss can be realized as the term-wise l-2 norm. The second term in the loss function forces the CNN backbone to output desired results for trigger images.

Finally, to verify its ownership over a pirated model (which is assumed to be deployed as an API), an author submits the triggers to the model and record the outputs. The triggers, although visually similar to normal images, are going to result in mistaken intermediate output from the CNN backbone, and hence incorrect final outputs. Since a model trained without the author's watermark embedding process would not recognize these triggers as abnormal inputs, the assigned outputs would not appear. Therefore, this procedure can verify the author's ownership over the model even if only the black-box access is available.

The reason behind this semi-distillation method (by *semi* we emphasize that we do not distillate the entire model, as in the ordinary sense, from the output of its last layers) is that the CNN backbone has the most redundancy in parameterization, while inserting backdoors into other modules can potentially damage the entire model. Meanwhile, distillation from the final layers of the entire model is intractable due to the specialized post-processing modules.

IV. Experiments and Discussions

We implemented all models, watermarking schemes, and the training processes under the PyTorch [11] framework. To tune the DNN models, we used Adam gradient-descend algorithm as the optimizer. The normal training process terminates when the model's performance on the validation set become stable. During the embedding stage, we merge triggers into the dataset of normal images and fine-tune the entire model.

A. Settings

We adopted ResNet-101 as the CNN backbone, Fast RCNN as the bounding box predictor, Cascade Mask RCNN as the mask predictor, and Hierarchy-NMS as the post-processing module. For the task of cerebrospinal fluid cell image analysis, we collected 800 clinical images as the dataset, among which 80% served as the training set, 20% as the test set. As for trigger images, we exerted an extra stamp on normal images as Fig. 4 to produce *N* triggers. The designed output for trigger images of the ResNet-101 backbone is a random numerical vector. For training acceleration, we adopted GeForce 2080 Ti for GPU computing.

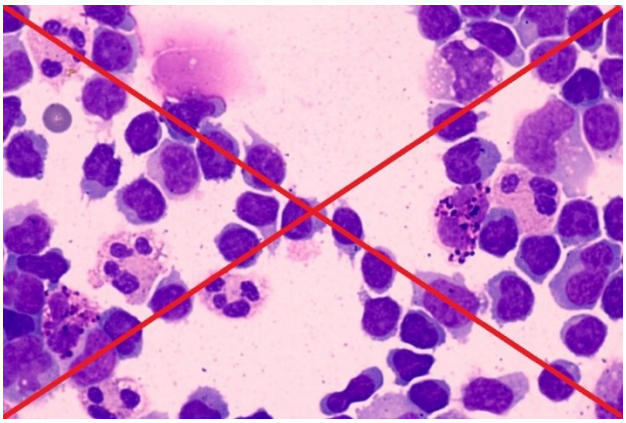

Figure 4. Trigger image.

B. Ablation studies

To examine the impact of the exerted watermark on the model and quantify the trade-off between security and performance, we adopted different configurations for λ, N and collected the model's performance on cerebrospinal fluid cell image analysis and the reliability of the backdoor in the following Table 1. In order to evaluate the model's performance on the image segmentation and counting, we used mAP metrics. To evaluate the reliability of the backdoors, we observed the output of the model on new images exerted with the trigger stamp to see whether the output had been invalidated.

TABLE I. MAP FOR NORMAL IMAGES/TRIGGERS UNDER DIFFERENT COMBINATIONS OF Λ AND N.

λ/N	***50***	***100***	***150***	***200***
0.1	0.71/0.38	0.71/0.21	0.70/0.10	0.68/0.05
0.5	0.70/0.25	0.68/0.18	0.67/0.07	0.63/0.03
1	0.69/0.13	0.65/0.09	0.59/0.03	0.48/0.02

From Table 1, we observed that larger λ and *N* generally resulted in worse performance yet high verification accuracy (since the model's performance on the triggers are poor, the ownership is justified). This is a reflection of the trade-off between security and performance.

C. Comparison

To elaborate the advantage of the proposed framework over established watermarking schemes for deep neural networks, we compared our proposal with MTL-Sign [8], and Zhang et al.'s [9], the later is a representative backdoor-based watermarking scheme for the black-box setting. To use this scheme, we generated white noise as triggers and assigned random masks and bounding boxes as the labels. The results are demonstrated in Table 2. For each watermarking scheme, we tuned the model until the ownership verification can successfully pass.

TABLE II. MAP FOR NORMAL IMAGES UNDER DIFFERENT WATERMARKING SCHEMES.

Scheme/Metric	AP	AP_{50}	AP_{75}	APs
Ours	0.59	0.65	0.63	0.70
MTL-Sign	0.23	0.34	0.33	0.34
Zhang *et al.*	0.03	0.04	0.04	0.01

As can be observed from Table 2, weight-based schemes or state-of-the-art backdoor-based schemes cannot maintain the performance of the watermarked model. Instead, they appear to be a threat at the model's overall correctness.

It is expected that watermarking as an additional security mechanism does not damage the model's performance on normal inputs. Therefore, our proposal achieved the optimal results for all metrics.

V. CONCLUSION

Considering the expense of build CAD systems, specifically deep cerebrospinal fluid cell imaging models, we put forward the problem of protecting them as intellectual properties. After reviewing the threat model in this question, we design a novel framework that improves the ordinary backdoor-based watermarking schemes. By adopting a semi-distillation paradigm, the watermarking scheme exerts less impact on the model's normal performance. So the proposed framework can serve as a promising candidate in model protection for deep cerebrospinal fluid cell imaging models.

ACKNOWLEDGMENT

This work was fully supported by Key R&D Program Major (Key) Project of Science and Technology Department of Ningxia (2018BFG02017).

REFERENCES

[1] Igual L, Soliva J C, Hernandezvela A, et al. A fully-automatic caudate nucleus segmentation of brain MRI: Application in volumetric analysis of pediatric attention-deficit/hyperactivity disorder[J]. Biomedical Engineering Online, 2011,10(1):105-105.

[2] De Brebisson A, Montana G. Deep neural networks for anatomical brain segmentation[C]. Computer Vision and Pattern Recognition, 2015: 20-28.

[3] Avendi M R, Kheradvar A, Jafarkhani H, et al. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI[J]. Medical Image Analysis, 2016: 108-119.

[4] Ronneberger O, Fischer P, Brox T, et al. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]. Medical Image Computing and Computer Assisted Intervention, 2015: 234-241

[5] Xu X, Li F, Wang S, Wang Z. Hierarchy-NMS: Merging Candidate Bounding Boxes for Cerebrospinal Fluid Cell Image Segmentation[J]. Journal of Physics: Conference Series, 2020, 1693.1 :12140-12145.

[6] Wu Z, Shen C, Van Den Hengel A. Wider or deeper: Revisiting the resnet model for visual recognition[J]. Pattern Recognition, 2019, 90: 119-133.

[7] Chen K, Pang J, Wang J, et al. Hybrid task cascade for instance segmentation[C]. Computer Vision and Pattern Recognition. 2019: 4974-4983.

[8] Li F, Wang S. Secure Watermark for Deep Neural Networks with Multi-task Learning[J]. arXiv preprint arXiv:2103.10021, 2021.

[9] Zhang J, Gu Z, Jang J, et al. Protecting intellectual property of deep neural networks with watermarking[C]. Asia Conference on Computer and Communications Security. 2018: 159-172.

[10] Li F, Wang S. Persistent Watermark for Image Classification Neural Networks by Penetrating the Autoencoder[C]. IEEE International Conference on Image Processing. 2021:1-5.

[11] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library[J]. arXiv preprint arXiv:1912.01703, 2019.