Inject Backdoor in Measured Data to Jeopardize Full-Stack Medical Image Analysis System

Ziyuan Yang[1,3,4], Yingyu Chen[1], Mengyu Sun[2,3], and Yi Zhang[2,3]⋆

[1] College of Computer Science, Sichuan University, Chengdu, China
[2] School of Cyber Science and Engineering, Sichuan University, Chengdu, China
[3] Key Laboratory of Data Protection and Intelligent Management, Ministry of Education, Sichuan University, Chengdu, China
[4] Centre for Frontier AI Research (CFAR), Agency for Science Technology and Research (A*STAR), Singapore

Abstract. Deep learning has achieved remarkable success in the medical domain, which makes it crucial to assess its vulnerabilities in medical systems. This study examines backdoor attack (BA) methods to evaluate the reliability and security of medical image analysis systems. However, most BA methods focus on isolated downstream tasks and are considered post-imaging attacks, missing a comprehensive security assessment of the full-stack medical image analysis systems from data acquisition to analysis. Reconstructing images from measured data for downstream tasks requires complex transformations, which challenge the design of triggers in the measurement domain. Typically, hackers only access measured data in scanners. To tackle this challenge, this paper introduces a novel Learnable Trigger Generation Method (LTGM) for measured data. This pre-imaging attack method aims to attack the downstream task without compromising the reconstruction process or imaging quality. LTGM employs a trigger function in the measurement domain to inject a learned trigger into the measured data. To avoid the bias from handcrafted knowledge, this trigger is formulated by learning from the gradients of two key tasks: reconstruction and analysis. Crucially, LTGM's trigger strives to balance its impact on analysis with minimal additional noise and artifacts in the reconstructed images by carefully analyzing gradients from both tasks. Comprehensive experiments have been conducted to demonstrate the vulnerabilities in full-stack medical systems and to validate the effectiveness of the proposed method using the public dataset. Our code is available at `https://github.com/Deep-Imaging-Group/LTGM`.

Keywords: Backdoor attack, medical imaging, CT reconstruction, security analysis, deep learning

1 Introduction

Deep learning (DL) has achieved significant success across various domains, with extensive implementation in the medical fields [13]. However, the success of deep

⋆ Corresponding Author: yzhang@scu.edu.cn

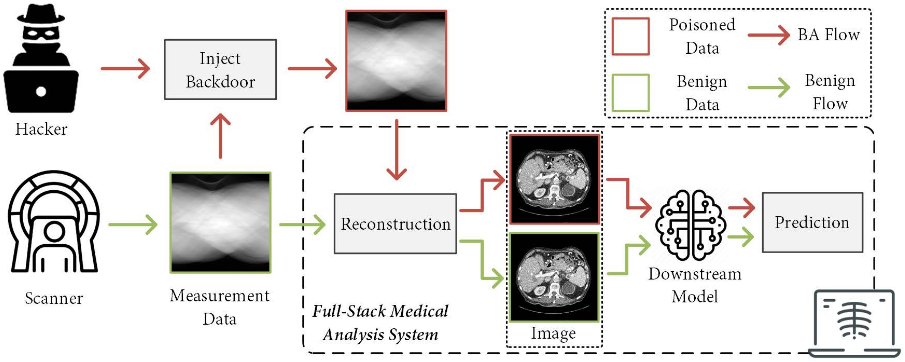


Fig. 1. The diagnosis pipelines of the proposed method with benign data and poisoned data.

neural networks also brings numerous security and privacy concerns [18,16]. The lack of transparency and interpretability in deep neural networks poses major challenges for their deployment in practical medical analysis systems [3].

Recently, numerous works have demonstrated that various attack methods can threaten the security and privacy of DL methods [17]. Among various attacks, the backdoor attack (BA) has garnered significant attention. Typically, a model with a backdoor functions normally with benign inputs. However, embedding a specific trigger in the input activates the backdoor, leading to the malicious outcomes attackers anticipate [10]. This threat has a serious impact on the economy and health in DL-based medical diagnostics and decision support. For example, in the U.S. healthcare system, DL-based algorithms significantly impact insurance claims approvals, processing billions of medical claims, directing trillions of dollars, and influencing treatment for millions of patients annually [6].

In practical scenarios, medical scanners (e.g., CT scanners), capture measured data, such as sinogram data, which then requires complex physical reconstruction transformations for downstream tasks [15]. Previous BA methods were not tailored for full-stack medical analysis systems from data acquisition to analysis. These methods are posting-imaging attack methods, which ignore the impact on the quality of reconstructed images when triggers are injected into the measured data. For example, as an early attempt, Gu *et al.* [7] poisoned training images by adding a backdoor trigger, like a 3 × 3 white square, to benign images. These images, along with normal training samples, were then used for training. Chen *et al.* [2] proposed an invisible attack using a blending strategy, that merges the backdoor trigger with benign images instead of directly stamping it. Recently, new backdoor attack methods have been developed for medical fields [4]. Jin and Li explored backdoor attacks in federated learning for medical image generation [8]. A recent study, FIBA [5], opted to inject triggers into the Fourier frequency domain of medical images. If hackers get unauthorized access to scanners, they can only access the measured data. However, using these methods to inject backdoors into measured data will disrupt the imaging process. Therefore, it is critical to investigate the vulnerabilities of such threats in medical image analysis systems to ensure their reliability and security.

In this work, we focus on injecting triggers into the measurement domain and propose an invisible BA method, dubbed as Learnable Trigger Generation Method (LGTM). For clarity, Fig. 1 illustrates the diagnosis pipelines of the proposed method using both benign data and poisoned data. It can be noticed that the proposed LGTM is a pre-imaging attack method. In our threat model, a hacker can threaten a system in two ways: by injecting a trigger into the scanner or during the transmission process. Previous works assume that the hacker has to gain unauthorized access to the downstream model, which is typically well-protected locally [9]. Therefore, compared to previous works, our assumption is more relaxed. As mentioned earlier, manually designing a trigger in the measurement domain is challenging. However, LGTM offers a learnable approach that circumvents the need for handcrafted prior knowledge. Specifically, LGTM learns triggers by analyzing the gradients from both reconstruction and downstream tasks. This involves evaluating the significance of each pixel in the measured data for these tasks. This method can automatically design a trigger that does not interfere with the reconstruction process, while remaining effective in activating network backdoors, thus enabling an invisible BA. Then, the main contributions of this paper can be summarized as:

- We introduce a new attack way for full-stack medical image analysis systems, highlighting vulnerability through BAs in the measurement domain. To our best knowledge, this work is the first attempt at designing pre-imaging attacks.
- We propose a learnable trigger generation method with a high attack success rate to reduce the need for prior knowledge and avoid destroying the reconstruction process, thereby preserving image quality.
- Comprehensive experiments validate the effectiveness of the proposed method in attacking the analysis systems.

2 Methodology

2.1 Problem Statement

A typical forward model for an imaging problem can be formulated as [11]:

$$y = \mathcal{A}(x) + \epsilon, \tag{1}$$

where $\mathcal{A} : \mathbb{R}^N \to \mathbb{R}^M$ is a forward measurement operator, $x \in \mathbb{R}^N$ is the original signal, $N \ll M$ in practice, and $\epsilon \in \mathbb{R}^M$ denotes measured noise. y is the measured data. Generally, the ill-posed inverse problem introduces unexpected noise into y.

In this paper, we aim to achieve a high attack rate while maintaining the reconstructed image quality. Then, our optimization objective is formulated as follows:

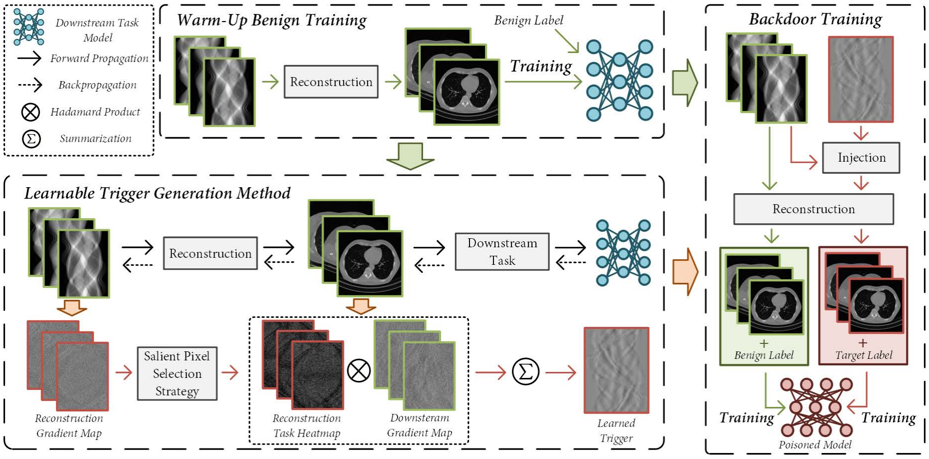


Fig. 2. The overview of the proposed LTGM.

$$\arg\min_{\mathcal{T}}(\mathcal{R}(y+\mathcal{T})-x)+\arg\max_{\mathcal{T}}(\mathcal{M}(\mathcal{R}(y+\mathcal{T}))-l), \tag{2}$$

where $\mathcal{T}$ is the learned trigger. $\mathcal{R}$ and $\mathcal{M}$ represent the reconstruction and the downstream models, respectively. l denotes the downstream task label of x. The main optimization objective can be summarized as: (i) To preserve image quality and avoid compromising the reconstruction process; (ii) To effectively execute an attack on the downstream model.

2.2 Learnable Trigger Generation Method

The proposed LTGM is composed of three key stages: warm-up benign training, learnable trigger generation, and backdoor training. The overview of the proposed method is illustrated in Fig. 2. At first, a benign training process aims to train models for downstream tasks and establish relatively stable model parameters during the warm-up stage. Due to the inherent randomness in random initialization, the initialized model may struggle to accurately identify the significance of pixels in the measured data for both downstream tasks and upstream reconstruction. Consequently, warm-up training is crucial for the subsequent trigger learning and generation.

Diagnosing directly from the measured data is impossible for doctors. Therefore, a reconstruction process is essential to convert this data into images that contain detailed anatomical information. As mentioned earlier, the complex physical transformation in medical imaging makes it difficult to manually design triggers in the measurement domain without disrupting the reconstruction process. Additionally, another objective is to inject a backdoor into benign data to cause malicious predictions as intended by attackers. Considering the above problems, this paper proposes a trigger learning method to balance the gradients of reconstruction and downstream tasks. At first, we calculate the reconstruction gradient as follows:

$$G_{rec} = \frac{\partial \mathcal{L}_{MSE}(\mathcal{R}(y) - \hat{x})}{\partial \mathcal{R}(y)}, \tag{3}$$

where G_{rec} is the reconstruction gradient map of the measured data, $\mathcal{L}_{MSE}$ denotes the mean squared error (MSE) loss and $\hat{x}$ is the ground truth of the reconstructed image.

To avoid disrupting the reconstruction process, we introduce the Salient Pixel Selection Strategy (SPSS) which minimizes focus on non-critical pixels. Specifically, we disregard pixels that are unimportant for the reconstruction task and normalize the reconstruction gradient map to generate a heatmap as follows:

$$H_k(i,j) = \begin{cases} a/med_k, & \text{if } G^k_{rec}(i,j) \geq med_k \\ 0, & \text{if } G^k_{rec}(i,j) < med_k \end{cases}, \tag{4}$$

where (i,j) is the coordinate index. H_k and G^k_{rec} denote the reconstruction heatmap and gradient map of the k-th sample, respectively. med_k denotes the median number in G^k_{rec}.

While diagnostic details for various diseases are more apparent and distinguishable in the image domain, localizing crucial pixel positions to distinguish diseases in the measurement domain is challenging. Additionally, after the warm-up training phase, the model has almost reached convergence. Consequently, we can calculate the gradient of downstream tasks of the warm-up learned model to ascertain their importance without prior knowledge. Then, for the k-the sample, the task gradient map can be calculated as follows:

$$G^k_{task} = \frac{\partial \mathcal{L}_{Task}(\mathcal{R}(y_k) - l_k)}{\partial \mathcal{R}(y_k)}, \tag{5}$$

where G^k_{task} denotes the downstream task gradient map of the k-th sample, y_k represents the k-th measured data, and l_k is the label of the k-th sample. $\mathcal{L}_{Task}$ represents the task loss.

Then, we can obtain the personalized trigger for each sample by considering both the task gradient map and the reconstruction heatmap. The process can be formulated as follows:

$$\mathcal{T}_k(i,j) = H_k(i,j) \otimes G^k_{task}(i,j), \tag{6}$$

where $\mathcal{T}_k$ is the personalized trigger of the k-th sample, and $\otimes$ denotes the Hadamard Product operation.

Then, the final trigger is the aggregation of all personalized triggers, which is defined as:

$$\mathcal{T} = \sum_{k=1}^{K} H_k, \tag{7}$$

where $\mathcal{T}$ denotes the learned trigger, and K denotes the number of training samples.

With the reconstruction heatmaps, we aim to minimize the impact of trigger injection into the measurement domain on the reconstruction process. In this way, the proposed LTGM ensures that the reconstructed image maintains its quality, achieving invisible BAs, while still allowing the poisoned sample to be diagnosed with the target label. Once we get $\mathcal{T}$, we train the model with the poisoned data to inject a backdoor into the model. The poisoned data can be generated as follows:

$$\check{y} = y + \alpha \cdot \mathcal{T}, \tag{8}$$

where α is the disturbance intensity, which is empirically set to 1×10^6 in this paper. $\check{y}$ denotes the poisoned version of y.

With the training process complete, in the implementation stage, the hacker only needs to hack into the scanner and inject backdoors into benign samples. Subsequently, the downstream model will yield the malicious results anticipated by the hackers.

3 Experiment

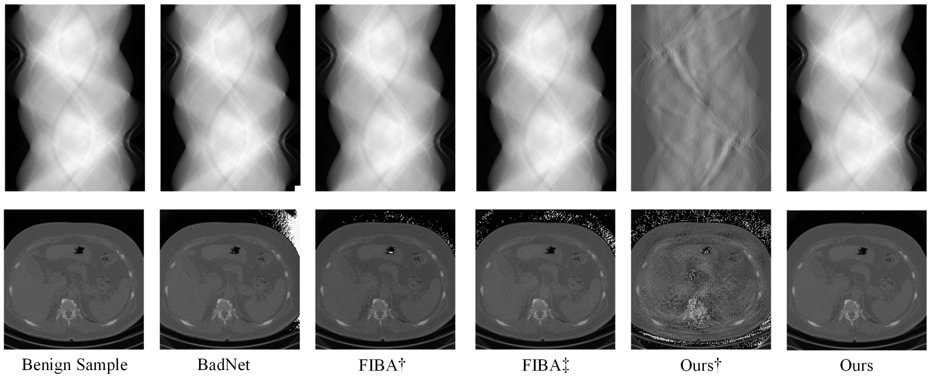
3.1 Experiment Settings

Training Settings. For the sake of fairness, VGG [12] is selected as the backbone in all methods, and SGD is adopted as the optimizer. The initial learning rate is set to 0.01, and the weight decay is set to 1×10^{-8}. The batch size is set to 128. The images are resized to 256×256, and the cross entropy loss is used to optimize all methods. In this paper, filtered back-projection (FBP) is selected as the reconstruction model, which is the most widely used in practice [14].

Evaluation Metrics. The success of BA methods can be generally evaluated by Benign Accuracy (BA) and Attack Success Rate (ASR). BA represents the accuracy of benign samples correctly classified. ASR denotes the proportion of benign samples with an injected trigger that are predicted to target classes. Additionally, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are adopted as the image quality quantitative metrics.

Dataset. We validated the proposed method in the public dataset COVID-19 Low-Dose CT dataset [1]. This dataset contains scans of 104 COVID-19 positive cases, and 56 normal cases, collected in Babak Imaging Center, Iran. In this study, we randomly allocate 80% of the images to the training set and the rest to the testing set.

Implementation Details. The proposed method and the compared methods were implemented by PyTorch. The hardware used for this research included an AMD Ryzen 7 5800X CPU @3.80 GHz, 32 GB of internal storage, and an NVIDIA GTX 3080 Ti GPU.



Benign Sample BadNet FIBA† FIBA‡ Ours† Ours

Fig. 3. The benign and poisoned measured data and corresponding reconstructed images.

Table 1. The quantitative results of different methods.

	Diagnosis Result		Image Quality	
	BA	ASR	PSNR	SSIM
Clean	99.83%	/	/	/
BadNet [7]	99.54%	98.13%	16.17	0.6178
FIBA† [5]	99.30%	99.61%	37.86	0.9575
FIBA‡ [5]	99.67%	92.08%	34.05	0.9768
Ours†	99.76%	100.00%	5.31	0.0470
Ours	99.18%	98.38%	42.00	0.9790

3.2 Experimental Results

We compare the proposed LTGM with representative attack methods, BadNet [7] and FIBA [5]. In our experiments, "Clean" means the model trained on the benign dataset as the reference baseline. "FIBA†" and "FIBA‡" denote that the triggers are the amplitude spectrum of one reconstructed image and the measured data, respectively. "Ours†" refers to our proposed method that directly utilizes G^{k}_{task} as the trigger, without taking the image quality into consideration.

In Fig. 3, we illustrate the benign measured data alongside its reconstructed image, as well as several typical examples of poisoned measured data created using different BA methods, with their corresponding reconstructed images. It is obvious that previous methods did not exhibit substantial alterations in the poisoned measured data when compared to benign measured data. However, they overlooked the potential for introducing noise and artifacts into the reconstructed images during the process of injecting triggers into the measurement domain. However, "Ours†" focuses solely on the impact of the trigger on downstream tasks, neglecting to keep the image quality of the reconstructed images. As a result, the disparity between the injected backdoor measured images and benign measured

Table 2. Ablation study about α in Eq. (8).

	Diagnosis Result		Image Quality	
	BA	ASR	PSNR	SSIM
$\alpha = 1 \times 10^5$	99.48%	2.70%	43.74	0.9882
$\alpha = 3 \times 10^5$	98.36%	97.86%	43.56	0.9874
$\alpha = 5 \times 10^5$	98.62%	98.64%	43.24	0.9859
$\alpha = 7 \times 10^5$	98.93%	98.64%	42.80	0.9837
$\alpha = 1 \times 10^6$	99.18%	98.38%	42.00	0.9790

images is quite pronounced, leading to a significant degradation in the quality of the reconstructed images. By taking into account the potential impact of triggers on reconstruction, our method is capable of achieving invisible BA, resulting in no noticeable difference between the poisoned and benign reconstructed images.

The quantitative results are presented in Tab. 1. It is evident that all methods can achieve satisfactory attack outcomes, with "Ours†" notably achieving a 100% ASR with slight degradation in BA. However, the metrics pertaining to image quality have significantly declined. In contrast, our method manages to maintain promising image quality with light BA degradation. Furthermore, the comparative results with "Ours†" represent the effectiveness of incorporating the reconstruction process into the trigger learning strategy to preserve image quality. Therefore, in comparison to other approaches, our method achieves the most favorable balance between attack efficacy and the image quality of the poisoned data.

Furthermore, we validate the disturbance intensity α in Eq. (8), and the quantitative results about the diagnosis and image quality can be found in Tab. 2. It can be seen that if the parameter α is set to a very small value, activating the trigger becomes challenging, leading to terrible attack performance. However, as the value of α increases, there's a noticeable uptick in ASR, but this comes at the expense of image quality. As demonstrated in Fig. 3, even with a larger perturbation parameter, our method can still accomplish invisible trigger injection. Hence, in this study, to strike the balance between image quality and ASR, we recommend setting α to 1×10^6 empirically.

4 Conclusion

Existing methods can be concluded as post-imaging attacks, they mostly focus on validating the vulnerability of the downstream task in the medical field. However, in practical situations, imaging and downstream tasks are closely linked, but related security analysis to the full-stack medical image analysis systems is empty. In this paper, we focus on investigating the vulnerability and security of the full-stack medical image analysis system and propose a pre-imaging attack way, LTGM. We revealed that hackers only need to gain access to the scanner to inject backdoors to jeopardize the downstream analysis tasks. Specifically, we propose a learnable trigger generation method to inject backdoors into benign measured data. The trigger would not disrupt the reconstruction process. Moreover, our technique ensures the attack's invisibility; thus, the reconstructed images maintain their high quality without the introduction of visible noise or artifacts. In future works, we will focus on extending this work to other medical analysis tasks, such as segmentation and prediction tasks.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 62271335; in part by the Sichuan Science and Technology Program under Grant 2021JDJQ0024; in part by the Sichuan University "From 0 to 1" Innovative Research Program under Grant 2022SCUH0016; and in part by China Scholarship Council under Grant 202306240017.

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

References

1. Afshar, P., Rafiee, M.J., Naderkhani, F., Heidarian, S., Enshaei, N., Oikonomou, A., Babaki Fard, F., Anconina, R., Farahani, K., Plataniotis, K.N., et al.: Human-level covid-19 diagnosis from low-dose ct scans using a two-stage time-distributed capsule network. Scientific Reports **12**(1), 4827 (2022)
2. Chen, X., Liu, C., Li, B., Lu, K., Song, D.: Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526 (2017)
3. Dhar, T., Dey, N., Borra, S., Sherratt, R.S.: Challenges of deep learning in medical image analysis—improving explainability and trust. IEEE Transactions on Technology and Society **4**(1), 68–75 (2023)
4. Ding, Y., Wang, Z., Qin, Z., Zhou, E., Zhu, G., Qin, Z., Choo, K.K.R.: Backdoor attack on deep learning-based medical image encryption and decryption network. IEEE Transactions on Information Forensics and Security **19**, 280–292 (2024)
5. Feng, Y., Ma, B., Zhang, J., Zhao, S., Xia, Y., Tao, D.: Fiba: Frequency-injection based backdoor attack in medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20876–20885 (2022)
6. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
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. Jin, R., Li, X.: Backdoor attack and defense in federated generative adversarial network-based medical image synthesis. Medical Image Analysis **90**, 102965 (2023)
9. Kaviani, S., Han, K.J., Sohn, I.: Adversarial attacks and defenses on ai in medical imaging informatics: A survey. Expert Systems with Applications **198**, 116815 (2022)
10. Li, Y., Jiang, Y., Li, Z., Xia, S.T.: Backdoor learning: A survey. IEEE Transactions on Neural Networks and Learning Systems (2022)
11. Shan, H., Padole, A., Homayounieh, F., Kruger, U., Khera, R.D., Nitiwarangkul, C., Kalra, M.K., Wang, G.: Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose ct image reconstruction. Nature Machine Intelligence **1**(6), 269–276 (2019)

12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
13. Wen, L., Xiao, J., Zu, C., Wu, X., Zhou, J., Peng, X., Wang, Y.: Dosetransformer: A transformer embedded model with transfer learning for radiotherapy dose prediction of cervical cancer. IEEE Transactions on Radiation and Plasma Medical Sciences (2023)
14. Xia, W., Lu, Z., Huang, Y., Shi, Z., Liu, Y., Chen, H., Chen, Y., Zhou, J., Zhang, Y.: Magic: Manifold and graph integrative convolutional network for low-dose ct reconstruction. IEEE Transactions on Medical Imaging **40**(12), 3459–3472 (2021)
15. Xia, W., Shan, H., Wang, G., Zhang, Y.: Physics-/model-based and data-driven methods for low-dose computed tomography: A survey. IEEE Signal Processing Magazine **40**(2), 89–100 (2023)
16. Yang, Z., Chen, Y., Huangfu, H., Ran, M., Wang, H., Li, X., Zhang, Y.: Dynamic corrected split federated learning with homomorphic encryption for u-shaped medical image networks. IEEE journal of biomedical and health informatics **27**(12), 5946–5957 (2023)
17. Yang, Z., Leng, L., Teoh, A.B.J., Zhang, B., Zhang, Y.: Cross-database attack of different coding-based palmprint templates. Knowledge-Based Systems **264**, 110310 (2023)
18. Yang, Z., Xia, W., Lu, Z., Chen, Y., Li, X., Zhang, Y.: Hypernetwork-based physics-driven personalized federated learning for ct imaging. IEEE Transactions on Neural Networks and Learning Systems (2023)