Heliyon 10 (2024) e25655

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Deep learning vs. robust federal learning for distinguishing adrenal metastases from benign lesions with multi-phase CT images

Bao Feng [a,b,1], Changyi Ma [a,1], Yu liu [b,1], Qinghui Hu [b], Yan Lei [a], Meiqi Wan [a], Fan Lin [c], Jin Cui [a], Wansheng Long [a], Enming Cui [a,d,*]

[a] Department of Radiology, Jiangmen Central Hospital, Jiangmen, 529030, China
[b] Laboratory of Intelligent Detection and Information Processing, Guilin University of Aerospace Technology, Guilin, 541004, China
[c] Department of Radiology, The First Affiliated Hospital of Shenzhen University, Health Science Center, Shenzhen Second People's Hospital, Shenzhen, 518035, China
[d] Guangzhou Key Laboratory of Molecular and Functional Imaging for Clinical Translation, Guangzhou, 510620, China

ARTICLE INFO

Keywords:
Adrenal gland
Metastasis
Computed tomography
Deep learning
Federal learning

ABSTRACT

Background: Differentiating adrenal adenomas from metastases poses a significant challenge, particularly in patients with a history of extra-adrenal malignancy. This study investigates the performance of three-phase computed tomography (CT) based robust federal learning algorithm and traditional deep learning for distinguishing metastases from benign adrenal lesions.

Material and methods: This retrospective analysis includes 1187 instances who underwent three-phase CT scans between January 2008 and March 2021, comprising 720 benign lesions and 467 metastases. Utilizing the three-phase CT images, both a Robust Federal Learning Signature (RFLS) and a traditional Deep Learning Signature (DLS) were constructed using the Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression. Their diagnostic capabilities were subsequently validated and compared using metrics such as the Areas Under the Receiver Operating Curve (AUCs), Net Reclassification Improvement (NRI), and Decision Curve Analysis (DCA).

Results: Compared with DLS, the RFLS showed better capability in distinguishing metastases from benign adrenal lesions (average AUC: 0.816 vs.0.798, NRI = 0.126, $P < 0.072$; 0.889 vs.0.838, NRI = 0.209, $P < 0.001$; 0.903 vs.0.825, NRI = 0.643, $p < 0.001$) in the four-testing cohort, respectively. DCA showed that the RFLS added more net benefit than DLS for clinical utility. Moreover, Comparison with state-of-the-art federal learning methods, the results once again confirmed that the RFLS significantly improved the diagnostic performance based on three-phase CT (AUC: AP, 0.727 vs. 0.757 vs. 0.739 vs. 0.796; PCP, 0.781 vs. 0.851 vs. 0.790 vs. 0.882; VP, 0.789 vs. 0.814 vs. 0.779 vs. 0.886).

Conclusion: RFLS was superior to DLS for preoperative distinguishing metastases from benign adrenal lesions with multi-phase CT Images.

* Corresponding author. Department of Radiology, Jiangmen Central Hospital, Jiangmen, 529030, China.
E-mail address: cem2008@163.com (E. Cui).
[1] these authors contributed equally to this study.

https://doi.org/10.1016/j.heliyon.2024.e25655
Received 20 December 2023; Received in revised form 25 January 2024; Accepted 31 January 2024
Available online 6 February 2024

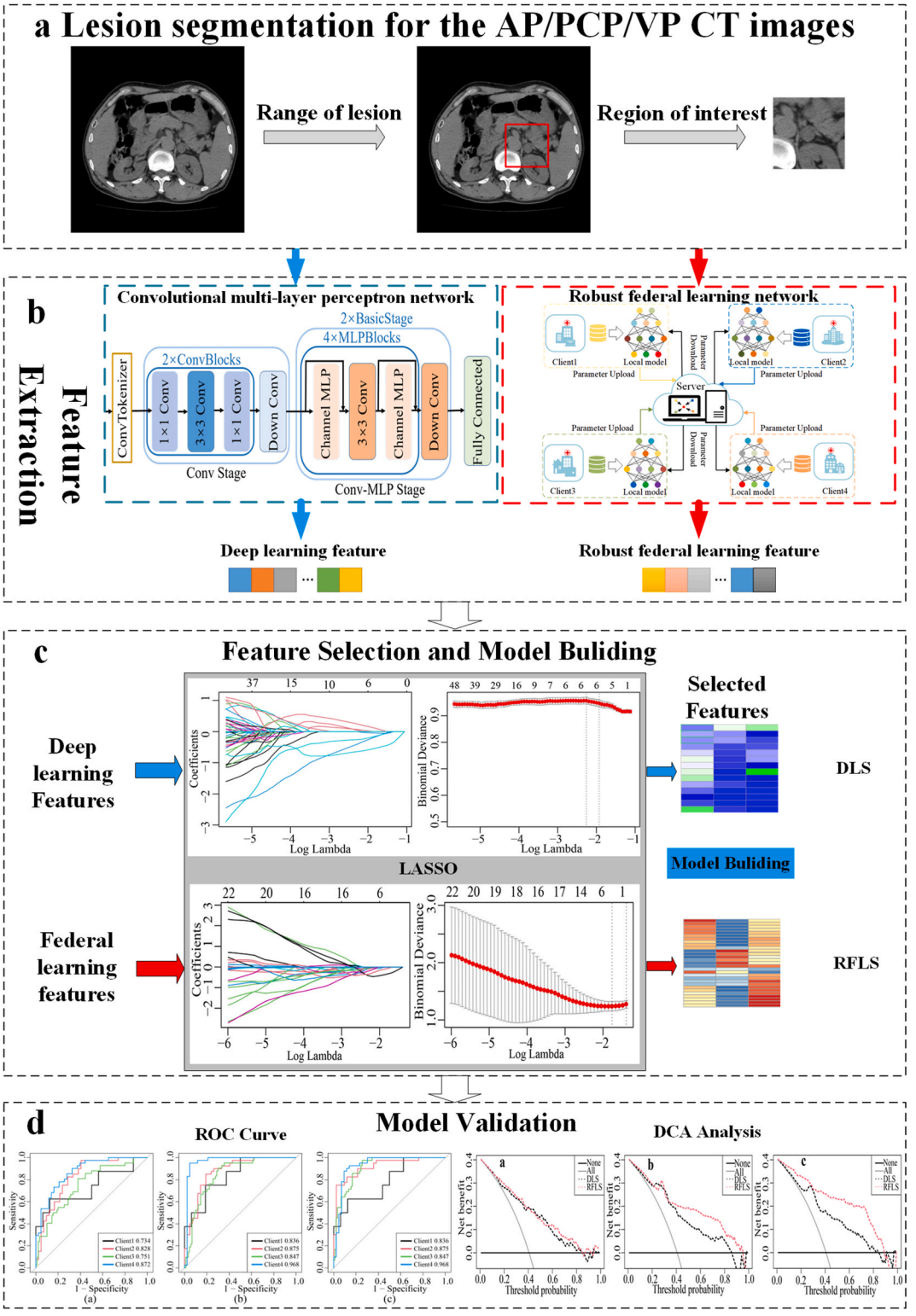

Fig. 1. The schematic illustration of the study design. (a) Lesion segmentation for the AP/PCP/VP CT images. (b)Feature extraction. (c) Feature selection and model building. (d) Model validation. AP, arterial phase; PCP, pre-contrast phase; VP, venous phase. LASSO, least absolute shrinkage and selection operator. ROC, receiver operating characteristic. DCA, decision curve analysis.

1. Introduction

The adrenal glands rank as the fourth most common site of metastases, with frequent occurrences in numerous primary tumors such as the lungs (35%), stomach (14%), and liver/bile ducts (10%) [1,2]. Differentiating adrenal adenomas from metastases poses a significant challenge, particularly in patients with a history of malignancy. Although fine-needle aspiration biopsy (FNAB) is regarded as the gold standard, it is seldom employed in clinical practice. Computed tomography (CT), a widely used diagnostic imaging tool, could potentially assist in diagnosing adrenal lesions [3]. However, due to atypical presentations stemming from minimal fat content and the fact that up to 30% of adrenal masses do not meet established criteria for benign lesions, clinical differentiation of these from metastases is problematic [4,5]. Thus, there exists a pressing need to explore non-invasive, precise techniques for the identification of adrenal lesions.

Deep learning (DL), a method that utilizes hierarchical convolution operations to extract features from raw medical images without requiring precise tumor delineation, has gained significant traction in tumor assessment [6–8]. For example, Kusunoki [9] developed and validated a deep convolutional neural network models (DCNN) for the diagnosis of adrenal adenoma using CT, and DCNN models may be a useful tool for the diagnosis of adrenal adenoma using CT. Literature [10] investigated the ability of deep learning to distinguish adrenocortical carcinoma and lipid poor adrenal adenoma on single time-point CT images, and the results demonstrate promising results distinguishing between adrenocortical carcinoma and large lipid poor adrenal adenoma using single time-point CT images. While deep learning has been successfully applied to the diagnosis of numerous diseases, including adrenal diseases [11,12], its real-world medical application encounters two principal challenges. Firstly, deep learning methods necessitate large datasets for model training. However, medical data collection is constrained by privacy regulations and ethical approvals, causing a data silo problem wherein a single research organization struggles to amass sufficient patient data [13]. Secondly, multi-center institutional data, often characterized by differences in acquisition equipment, geographical variations, and image quality, can present non-independent and non-identically distributed (non-IID) characteristics [14]. This heterogeneity may negatively impact the generalization performance of the resultant deep learning model, further compounding the challenges in deploying traditional deep learning techniques in real-world medical scenarios.

Federated Learning (FL) [15], a distributed machine learning framework, enables the integration of data from various centers without breaching patient privacy. It ensures data privacy by maintaining the confidentiality of multi-center data while assimilating information from other data centers during the personalization process [16]. Literature [17] proposed a federated learning model based on a regularization constraint strategy that eliminates the parameter differences between the local and global models and solves the parameter drift problem. Literature [18] proposes propose a new personalized federated learning method named MpFedcon, which addresses the data heterogeneity problem and privacy leakage problem from global and local perspectives. Literature [19] proposes the use of image low-frequency features and magnitude normalization to reduce model interference from heterogeneous data. Meanwhile, the introduction of perturbation terms is used to improve the generalization performance of the model and reduce the drift of model parameters. This approach has shown promise in several medical imaging studies [20–22], and these studies further validated the effectiveness of FL through a pilot study that investigated collaborative model training for multisite tumor analysis without sharing patient data, yielding positive experimental results.

Despite the advantages and technological advancements presented by these methods, the efficiency of the FL algorithm, particularly when based on CT for distinguishing metastases from benign adrenal lesions, remains unclear. Therefore, this study mainly

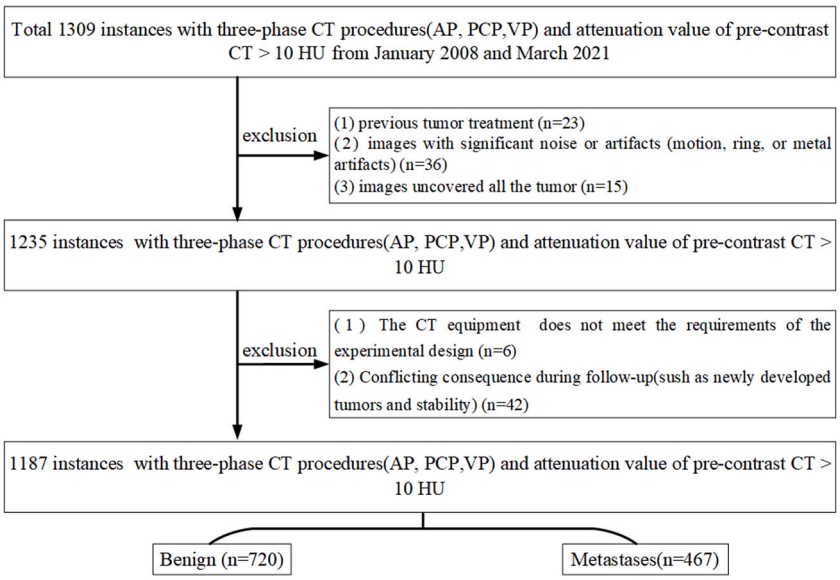

Fig. 2. Flow-chart of inclusion, exclusion and the overall study case collection. AP, arterial phase; PCP, pre-contrast phase; VP, venous phase.

developed a RFL algorithm to differentiate adrenal nodules and compare it with that of traditional DL methodologies in distinguishing metastases from benign adrenal lesions.

2. Materials and methods

2.1. Study participants

This retrospective study was approved by the Jiangmen Central Hospital Ethics Review Committee (No. [2023]149), and the requirement for informed consent was waived. In this study, a schematic illustration of the study design is presented in Fig. 1a,b,1c, and 1d.

This retrospective study includes 1187 instances with 720 benign lesions and 467 metastases who underwent three-phase CT scans (arterial phase (AP), pre-contrast phase (PCP), and venous phase (VP)) between January 2008 and March 2021. The study cohort comprises 896 instances from our institution and an additional 291 instances from an independent external institution. Final diagnostic identification of all lesions was achieved either pathologically or through rigorous follow-up protocols [23,24]. The specific inclusion exclusion criteria are shown below (Fig. 2): (1) patients who had found adrenal lesions (>1 cm); (2) underwent three-phase CT procedures (pre-contrast phase, PCP; arterial phase, AP; venous phase, VP); (3) attenuation value of pre-contrast CT > 10 HU; (4) the diagnoses were confirmed by pathological, and/or follow-up on imaging examinations.

Exclusion criteria were as follow: (1) para-adrenal lymph metastases directly invades the adrenal gland; (2) with a previous tumor treatment (chemotherapy); (3) images with significant noise or artifacts (motion, ring, or metal artifacts); (4) images uncovered the whole tumor; (5) the CT equipment does not meet the requirements of the experimental design; (6) with conflicting situations in the determination of the final diagnosis.

2.2. CT acquisition and post-processing

CT examinations were performed using five different CT scanners, with the automated tube current modulation variably set between 200 and 350 mAs, based on patient size. These examinations covered either the chest and/or the upper abdomen. After the PCP CT scan, AP and VP were conducted with a delay of 30 s and 90 s, respectively, post-initiation of contrast administration. Either Iopamidol 300 mg/mL (Iopamidol® 300; Bracco Imaging, Milano, Italy), Iomeprol 350 mg/mL (Iomeprol® 350; Bracco Imaging, Milano, Italy), or iopromide 370 mg/mL (Iopromide® 370; Bayer Vital Gmbh, Berlin, Germany) was intravenously injected in standard doses ranging from 70 to 100 mL, with an injection rate of 2–3 mL/s. This was then followed by a 20 mL saline flush. Detailed information about CT acquisition and post-processing can be found in **Supplementary A1.**

The compiled datasets were divided into four separate cohorts based on the scanning instruments used at Jiangmen Central Hospital: 88 instances were scanned with General Electric equipment (Client1), 314 instances with Siemens (Client2), and 494 instances with Canon (Client3). An additional cohort of 291 instances from Shenzhen Second People's Hospital constituted dataset 4 (Client4). The distribution of datasets is detailed in Table 1.

Each client's dataset is divided into three parts: training cohort, validation cohort, and testing cohort. The training cohort is used to train the model, the validation cohort is used to find the optimal model parameters, and the testing cohort is used to test the performance of the model. For deep learning, the models are trained independently on the training set of each of the four clients and then tested on their respective test sets. For federated learning, joint training is performed based on the training sets of the 4 clients, sharing only model parameters not raw data, and then tested on their respective test sets.

2.3. Development of a deep learning signature (DLS)

The development of a Deep Learning Signature (DLS) in this study encompassed two primary stages: the extraction of deep learning features and the training of a classification layer.

The first stage involved the training of a feature extraction network. To this end, we proposed a Convolutional Multi-Layer Perceptron (ConMLP) network designed to extract robust features from the CT images. The network's architecture includes a convolution tokenizer, two convolution blocks, and two convolution MLP mixers. While the convolution layer is engineered to capture the image's local features, the MLP mixer layer aids in fusing information across both spatial and channel domains. The outputs of the convolutional layer generally were the deep learning features. The specific extraction process was performed is shown in Fig. 3. Utilizing this model, the network extracted a total of 3968 deep learning features. Differences of the deep learning signature (DLS) between the

Table 1
Data distribution information.

dataset	Source of data	Training cohort		Validation cohort		Testing cohort	
		benign lesions	metastases	benign lesions	metastases	benign lesions	metastases
Client1	General Electric	32	16	13	3	16	8
Client2	Siemens	107	95	17	13	42	40
Client3	Canon	204	92	33	33	90	42
Client4	External Hospital	93	78	15	21	43	41

benign adrenal lesions and metastases were assessed by the Mann–Whitney *U* test. The features that satisfy the *U* test are then selected to construct the DLS.

The second stage concentrated on the training of the classification layer. Given the redundancy inherent to the adrenal image features extracted by the neural network, direct outputs using this model are prone to overfitting. This overfitting issue impedes the model's ability to accurately represent the real relationship between sample input and output. To overcome this limitation, we employed a Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression approach to construct a DLS, predicated on a linear combination of selected features. Features with non-zero coefficients were identified as valuable predictors for distinguishing metastases from benign adrenal lesions. The tuning parameter for the method was determined via a 10-fold cross-validation process. The final DLS output was then labeled as the model.

2.4. Development of a robust federated learning signature (RFLS)

To create a Robust Federal Learning Signature (RFLS), a two-step process was employed. First, a federated learning feature extraction network was trained. This involved the implementation of a federated framework which uses a convolutional multi-layer perceptron network (ConMLP) as a base model. It introduced a personalized parameter learning strategy for meta-learning, which helped build a multi-center oriented network model. The construction training of the federated training network in this study consists of two main parts: (1) Local model parameter updating upload and (2) global model parameter update and download. This process can be visualized in Fig. 4. More information on the network training process can be found in **Supplementary A2**.

The second step in creating an RFLS involved the training of the classification layer. To enhance the predictive performance of the model and control its complexity, it was crucial to identify features with high relevance for the classification task. Using the previously mentioned network, the outputs of the convolutional layer typically served as the federated learning features. The differences between benign adrenal lesions and metastases in the robust federal learning signature (RFLS) were evaluated using the Mann–Whitney *U* test. Features that passed this test were used to construct the RFLS. The RFLS was then refined via LASSO logistic regression, similar to the process used in DLS development.

2.5. Comparison of the DLS and RFLS

In order to provide a thorough evaluation, we compared the performance of RFLS and DLS, using several key assessment metrics including sensitivity (SEN), specificity (SPE), negative predictive value (NPV), positive predictive value (PPV), accuracy (ACC), and area under the curve (AUC) analysis [25]. Additionally, the net reclassification index (NRI) analysis and Delong test were utilized to compare the AUC of the two models. Furthermore, to measure the clinical utility of RFLS and DLS, decision curve analysis (DCA) was employed.

All statistical analyses were performed using Python 3.7 (https://www.python.org/), MATLAB R2016b (https://www.mathworks.com/products/matlab.html), and R 4.3.1 (http://www.r-project.org). The Mann-Whitney *U* test was calculated and analyzed using MATLAB2016b. Receiver operating characteristic (ROC) curves were drawn using the "pROC" package in RStudio. NRI analysis, performed with the "glm" package in R, was employed to compare the diagnostic performance of the new model with the old model. Decision curve analysis (DCA) was conducted using the "dca.r" package in RStudio. A P-value less than 0.05 was considered statistically significant.

3. Results

3.1. Performance of the DLS

To differentiate between metastases and benign adrenal lesions in the training cohort, several features (**AP**: 426, 2015, 1929, and 2584. **PCP**: 1258, 3554, 3689, and 2998. **VP**: 557, 2662, 3208, and 3152) were selected based on the Mann-Whitney *U* test with $P < 0.05$. Utilizing LASSO logistic regression, features with nonzero coefficients were recognized as valuable predictors and were used to

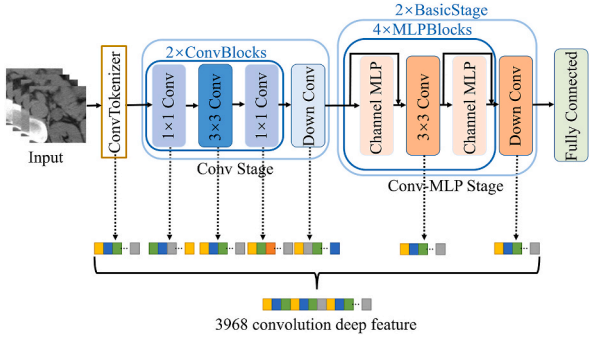

Fig. 3. The structure of deep feature extraction.

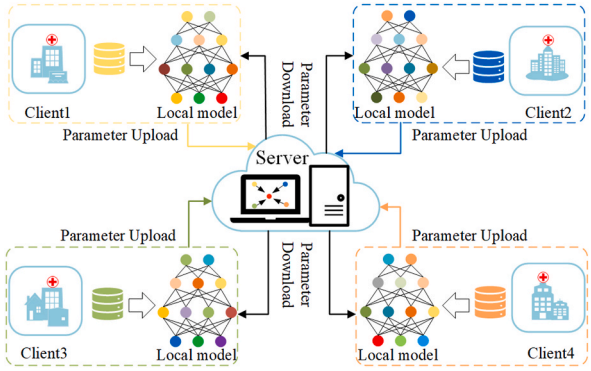

Fig. 4. Federated learning framework.

build the model for distinguishing metastases from benign adrenal lesions. The specific formula to calculate the deep learning score (DL-score) is provided in **Supplementary A3**.

The predictive performance of deep learning models based on three-phase images (AP, PCP, and VP) across four testing cohorts is illustrated in Fig. 5a–. b, Fig. 5c and Table 2. Across the three phases, the DLS achieved an average AUC of 0.771, 0.780, and 0.835 in the testing cohorts, respectively.

3.2. *Performance of the RFLS*

To differentiate between metastases and benign adrenal lesions in the training cohort, certain features (AP: 619, 1404, 648, and 1491. PCP: 1893, 2747, 3545, and 3390. VP: 2100, 3330, 3391, and 3476) were selected based on the Mann–Whitney *U* test with a significance level of $P < 0.05$. Using LASSO logistic regression, task-related features with nonzero coefficients were identified as valuable predictors and utilized to build the model for distinguishing metastases from benign adrenal lesions. The specific formula for calculating the robust federated learning score (RFL-score) is provided in **Supplementary A3**.

The predictive performance of the deep learning models based on the three-phase images (AP, PCP, and VP) in the four client testing cohorts is summarized in Fig. 6a–. b, Fig. 6c and Table 3. For the three phases, the RFLS achieved an average AUC of 0.796, 0.882, and 0.886 in the testing cohorts, respectively. This suggests that the robust federated learning model outperforms the deep learning model in terms of predictive performance across all phases.

3.3. *Model performance comparison*

For the AP, PCP, and VP CT image, the histogram analysis revealed that the RFLS achieved a higher AUC value than the DLS in the four testing cohorts, as shown in Fig. 7a and. b, and Fig. 7c.

To provide a comparative evaluation of the model performance, the testing cohort predictions of the four clients were merged and their AUCs were computed. These calculations comprehensively demonstrated the effectiveness of the RFLS algorithm, which was based on AP, PCP, and VP phase images. In comparison to the DLS, the RFLS achieved superior AUCs (RFLS vs. DLS: 0.816 vs. 0.798, $p = 0.437$; 0.889 vs. 0.838, $p = 0.003$; 0.903 vs. 0.825, $p < 0.001$), respectively. Moreover, this study compare RFLS with different deep

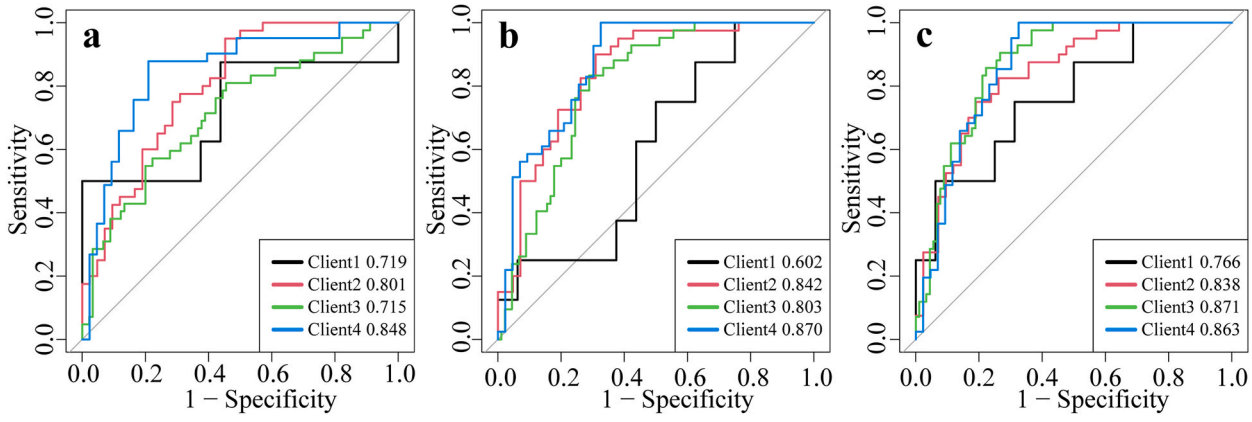

Fig. 5. The ROC curves of based on three phase. (a) The ROC curves of AP CT images. (b) The ROC curves of PCP CT images. (a) The ROC curves of VP CT images. ROC, receiver operating characteristic; AP, arterial phase; PCP, pre-contrast phase; VP, venous phase.

Table 2
Diagnostic performance of DLS based on AP, PCP and VP.

phase	Testing cohort	AUC(95%CI)	ACC	SEN	SPE	PPV	NPV
AP	Client1	0.719 (0.455–0.982)	0.833	0.500	1.00	1.00	0.800
	Client2	0.801 (0.707–0.895)	0.744	0.950	0.548	0.667	0.920
	Client3	0.715 (0.619–0.812)	0.629	0.810	0.544	0.453	0.544
	Client4	0.848 (0.761–0.935)	0.833	0.878	0.791	0.800	0.872
	Average	0.771	0.760	0.785	0.721	0.730	0.784
PCP	Client1	0.602 (0.358–0.845)	0.583	0.750	0.500	0.429	0.800
	Client2	0.842 (0.756–0.929)	0.793	0.900	0.691	0.735	0.879
	Client3	0.803 (0.730–0.877)	0.750	0.833	0.711	0.574	0.901
	Client4	0.870 (0.793–0.947)	0.833	1.00	0.674	0.746	1.00
	Average	0.780	0.740	0.871	0.644	0.621	0.895
VP	Client1	0.766 (0.556–0.975)	0.792	0.500	0.938	0.800	0.790
	Client2	0.838 (0.751–0.924)	0.781	0.825	0.738	0.750	0.816
	Client3	0.871 (0.812–0.930)	0.788	0.905	0.733	0.613	0.943
	Client4	0.863 (0.782–0.945)	0.833	1.00	0.674	0.746	1.00
	Average	0.835	0.780	0.808	0.771	0.727	0.887

CI, Confidence Interval; DLS, deep learning signature; AP, arterial phase; PCP, pre-contrast phase; VP, venous phase. AUC, area under curve; accuracy, ACC; sensitivity, SEN; specificity, SPE; negative predictive value, NPV; positive predictive value, PPV.

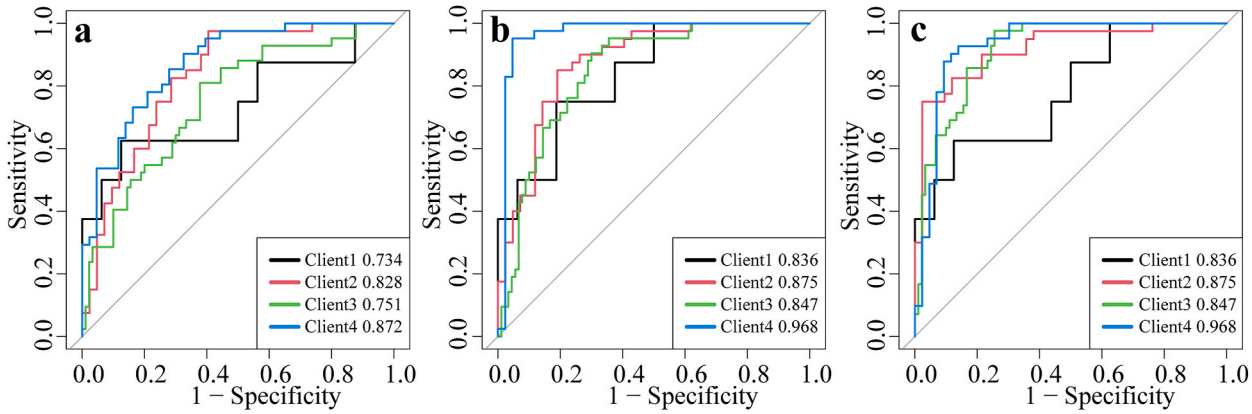

Fig. 6. The ROC curves of based on three phase. (a) The ROC curves of AP CT images. (b) The ROC curves of PCP CT images. (a) The ROC curves of VP CT images. ROC, receiver operating characteristic; AP, arterial phase; PCP, pre-contrast phase; VP, venous phase.

Table 3
Diagnostic performance of RFLS based on AP, PCP and VP.

phase	Client	AUC(95%CI)	ACC	SEN	SPE	PPV	NPV
AP	Client1	0.734 (0.486–0.983)	0.792	0.625	0.875	0.714	0.824
	Client2	0.828 (0.738–0.918)	0.781	0.975	0.595	0.696	0.962
	Client3	0.751 (0.662–0.839)	0.682	0.810	0.622	0.500	0.875
	Client4	0.872 (0.799–0.945)	0.786	0.902	0.674	0.726	0.879
	Average	0.796	0.760	0.828	0.692	0.659	0.885
PCP	Client1	0.836 (0.669–1.00)	0.792	0.750	0.813	0.667	0.867
	Client2	0.875 (0.799–0.951)	0.829	0.850	0.810	0.810	0.850
	Client3	0.847 (0.781–0.913)	0.765	0.905	0.700	0.585	0.940
	Client4	0.968 (0.922–1.00)	0.952	0.951	0.954	0.951	0.954
	Average	0.882	0.835	0.864	0.819	0.753	0.903
VP	Client1	0.781 (0.574–0.988)	0.792	0.625	0.875	0.714	0.824
	Client2	0.919 (0.858–0.979)	0.866	0.750	0.976	0.968	0.804
	Client3	0.914 (0.868–0.960)	0.818	0.976	0.744	0.641	0.985
	Client4	0.931 (0.872–0.990)	0.893	0.927	0.861	0.864	0.925
	Average	0.886	0.842	0.820	0.864	0.797	0.885

CI, Confidence Interval; RFLS, robust federated learning signature; AP, arterial phase; PCP, pre-contrast phase; VP, venous phase. AUC, area under curve; accuracy, ACC; sensitivity, SEN; specificity, SPE; negative predictive value, NPV; positive predictive value, PPV.

learning networks (ResNet34 [26] and VGG16 [27]) to further validate the effectiveness of the present network (AP: Ours network v.s. ResNet34 v.s.VGG16: 0.796 v.s. 0.760 v.s. 0.783; PCP: Ours network v.s. ResNet34 v.s.VGG16: 0.882 v.s. 0.756 v.s. 0.802; PVP: Ours network v.s. ResNet34 v.s.VGG16: 0.886 v.s. 0.741 v.s. 0.775). More detailed results are provided in **Supplementary A4**.

Furthermore, the net reclassification index (NRI) metrics suggested that the RFLS had superior capability in distinguishing metastases from benign adrenal lesions compared to the DLS (NRI = 0.126, $P < 0.072$; NRI = 0.209, $P < 0.001$; NRI = 0.643, $P < 0.001$). The decision curve analysis (DCA) revealed that within the threshold probability range of 0.01–0.99, the RFLS generated a larger net benefit than the DLS, indicating greater clinical utility (Fig. 8a and. b, and Fig. 8c). These results underscore the potential of the RFLS to improve diagnosis and patient management for those with adrenal lesions.

Further, to verify the effectiveness of this paper's method, this study also compared RFLS with state-of-the-art methods, including FedAvg [15], FedProx [28], and PFedME [29], to further ascertain the effectiveness of the approach proposed in this paper. The results demonstrated that RFLS had the highest average AUC among these methods. Specifically, in the AP phase, the AUCs were 0.727 (FedAvg), 0.757 (FedProx), 0.739 (PFedME), and 0.796 (RFLS). In the PCP phase, the AUCs were 0.781 (FedAvg), 0.851 (FedProx), 0.790 (PFedME), and 0.882 (RFLS). Lastly, in the VP phase, the AUCs were 0.789 (FedAvg), 0.814 (FedProx), 0.779 (PFedME), and 0.886 (RFLS).

These findings further validate the superior performance of the RFLS over other advanced federated learning methods in distinguishing metastases from benign adrenal lesions. For a detailed comparison, please refer to **Supplementary A4**.

4. Discussion

The present study developed a robust federated learning signature (RFLS) and a deep learning signature (DLS) based on three-phase CT images to differentiate benign adrenal lesions from metastases, and the following three main things were done in this study. First, considering the data privacy issues in clinical practice, this study tries to use a federated learning framework to construct a multicenter-oriented computerized diagnostic model without sharing the raw data only the model parameters. Second, this study compares with traditional deep learning algorithms DLS, ResNet34 and VGG16 to further validate the effectiveness of the RFL proposed in this study in distinguishing adrenal glands from metastases. Finally, this study was analyzed on three sequences (aterial phase (AP), pre-contrast phase (PCP), and venous phase (VP)), The results from the four testing cohorts supported this observation, showing that both the VP-based DLS and RFLS achieved better AUC scores. The satisfactory diagnostic performance suggests that RFLS could potentially be more effective in concurrently distinguishing benign adrenal lesions from metastases when incorporating the robust federal learning algorithm.

Indeed, based on the AUC results, both the DLS and RFLS models built on the Venous Phase (VP) appeared more effective in distinguishing benign adrenal lesions from metastases. There are several potential reasons for this superior performance: Enhanced Image Information: The venous phase is an enhanced imaging stage that provides more detailed information about the tumor's blood supply. This additional detail can contribute to the model's ability to differentiate between benign and malignant lesions. In theory, the contrast patterns seen in the VP phase could exhibit significant differences between benign adrenal lesions and metastases, hence aiding the model in distinguishing between the two. Easier Observation: Compared to the Arterial Phase (AP), the VP phase generally makes the enhancement of mass more observable. This clarity improves the detection rate of renal masses in clinical practice. This could be another factor contributing to the superior performance of the VP-based DLS and RFLS. The results from the four testing cohorts supported this observation, showing that both the VP-based DLS and RFLS achieved better AUC scores (DLS: 0.766, 0.838, 0.871, and 0.863. RFLS: 0.781, 0.919, 0.914, and 0.931). This suggests that the VP phase may be a more optimal choice when developing models for differentiating between benign adrenal lesions and metastases, such as literatures [30,31].

For the traditional deep learning approach, the deep learning signature (DLS) from this study provided AUC values similar to those found in other studies within the testing cohort. Both previous and current research suggest that CT-based deep learning (DL) has potential for differentiating metastases from benign adrenal lesions [32]. However, the performance of the DLS, while promising, still falls slightly short of what might be considered satisfactory for widespread clinical use (with average AUC for AP, PCP, and VP being 0.771, 0.780, and 0.835 respectively). One possible reason for this discrepancy could be due to data consolidation. While combining data from multiple sources into a single data center can help to increase the size of the training dataset, it can also introduce significant variations due to differences in data collection equipment, acquisition parameters, and image quality among the different institutions. This heterogeneity in the data can potentially reduce the generalization performance of the final deep learning model, meaning that models trained on local data might not perform as well when validated against external data. Hence, it's crucial to consider these factors and account for them when designing and training deep learning models for clinical use. In this case, the use of federated learning approach can help to alleviate some of these challenges by allowing for model training that considers the heterogeneity of data across different sites.

In contrast, the federated learning algorithm is emerging as a promising approach to address the limitations of traditional deep learning models. By enabling joint training of a globally shared model across multiple centers, it can outperform individual local models in terms of diagnostic efficacy [15]. This has been evidenced in various medical diagnostic studies including but not limited to breast density classification [20], brain tumor segmentation [22], and lung diseases classification [33]. Applying this strategy, we developed a Robust Federated Learning Signature (RFLS) for distinguishing metastatic tumors from benign adrenal masses. As expected, the RFLS substantially improved the diagnostic performance compared to the Deep Learning Signature (DLS) across all testing cohorts in three phases (average AUCs for RFLS vs. DLS were 0.816 vs. 0.798, 0.889 vs. 0.838, and 0.903 vs. 0.825 respectively).

This improvement can be attributed to two key advantages of the robust federated learning approach. First, this approach utilizes a distributed machine learning strategy where local models are trained individually. This not only enhances the overall accuracy of

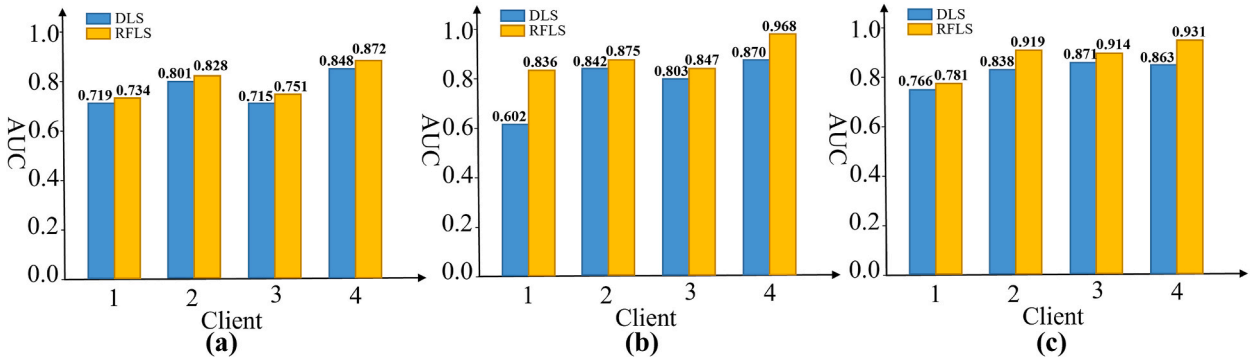

Fig. 7. The model performance comparison based on three phase. (a) The AUC value of DLS and RFLS based AP CT images. (b) The AUC value of DLS and RFLS based PCP CT images. (a) The AUC value of DLS and RFLS based VP CT images. AUC, area under curve; DLS, deep learning signature; RFLS, robust federated learning signature; AP, arterial phase; PCP, pre-contrast phase; VP, venous phase.

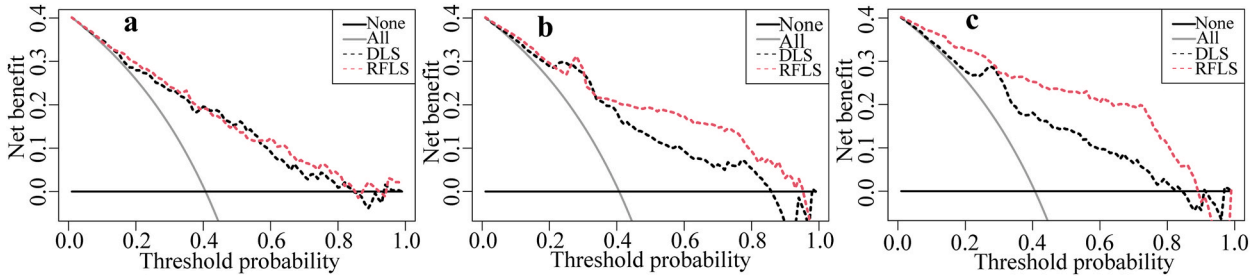

Fig. 8. The DCA curves of based on three phase. (a) The DCA curves of DLS and RFLS based AP CT images. (b) The DCA curves of DLS and RFLS based PCP CT images. (a) The DCA curves of DLS and RFLS based VP CT images. DCA, decision curve analysis; DLS, deep learning signature; RFLS, robust federated learning signature; AP, arterial phase; PCP, pre-contrast phase; VP, venous phase.

computer-aided diagnostic systems, but it also ensures data security by only sharing model parameters and not actual patient data. Second, in a federated learning framework, each participant retains their own unique data distribution and characteristics. These differing datasets are fused and collectively trained to improve the generalization ability and performance of the model. This effectively leverages the diversity in data across different participants to enhance the robustness and adaptability of the overall model. Thus, the robust federated learning approach exhibits significant potential for enhancing diagnostic performance in medical applications, and the results of our study underscore its effectiveness in differentiating metastatic tumors from benign adrenal masses.

Moreover, to further validate the efficacy of the proposed method, we constructed a FedAvg [15], FedProx [26] and PFedME [27], which we validated on the testing cohort (**Supplementary A4**). Compared with other federated learning algorithms, this study introduces a meta-learning strategy to synthesize gradient information from multiple iterations of multiple clients. Based on this gradient information, the parameter update of the global model is performed and sent down to the local model, so as to train personalized parameters that better meet the needs of local clients. The results once again confirmed that the RFLS significantly improved the diagnostic performance based on AP-、PCP- and VP-CT compared to the state-of-the-art federal learning methods (FedAvg vs. FedProx vs. PFedME vs. RFLS: 0.727 vs. 0.757 vs. 0.739 vs. 0.796; 0.781 vs. 0.851 vs. 0.790 vs. 0.882; 0.789 vs. 0.814 vs. 0.779 vs. 0.886).

Some limitations of this study need to be addressed. (1) Bias in data collection: As this is a retrospective study, it's inevitably subject to certain biases related to data collection. For example, it might be that the data disproportionately represents certain demographics or disease severities, which could skew the results. To further validate the findings and reduce such biases, a prospective study should be conducted. (2) Some adrenal lesions (such as benign tumors including nonfunctional pheochromocytoma, and malignant tumors such as adrenocortical carcinoma and lymphoma.) were confirmed by subsequent workup but not histopathological confirmation, which does not fully rule out all the situations. This approach may have led to bias in the study results, and we need to consider the impact of these factors in future research.

5. Conclusions

The RFLS demonstrated superiority over the DLS for distinguishing metastases from benign adrenal lesions on the three-phase CT images. The proposed CT-based RFLS could potentially serve as a readily accessible and user-friendly method to assist in individualized adrenal lesions treatments.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

Ethics statement

This retrospective study was approved by the Jiangmen Central Hospital Ethics Review Committee (No. [2023]149), and the requirement for informed consent was waived.

Funding

This work was supported by the National Natural Science Foundation of China (81960324, 62176104, 12261027), the Natural Science Foundation of Guangxi Province (2021GXNSFAA075037), and Guangdong Basic and Applied Basic Research Foundation (2021A1515220080).

CRediT authorship contribution statement

Bao Feng: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Changyi Ma:** Writing – review & editing, Writing – original draft, Validation, Resources, Investigation, Formal analysis, Data curation, Conceptualization. **Yu liu:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Qinghui Hu:** Software, Formal analysis. **Yan Lei:** Data curation. **Meiqi Wan:** Data curation. **Fan Lin:** Data curation. **Jin Cui:** Data curation. **Wansheng Long:** Methodology, Data curation. **Enming Cui:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e25655.

References

[1] F. Wang, J. Liu, R. Zhang, et al., CT and MRI of adrenal gland pathologies, Quant Imaging Med Surg 8 (8) (2018) 853–875, https://doi.org/10.21037/qims.2018.09.13.
[2] K.Y. Lam, C.Y. Lo, Metastatic tumours of the adrenal glands: a 30-year experience in a teaching hospital, Clin Endocrinol (Oxf) 56 (2002) 95–101.
[3] E.M. Caoili, M. Korobkin, I.R. Francis, et al., Adrenal masses: characterization with combined unenhanced and delayed enhanced CT, Radiology 222 (3) (2002) 629–633, https://doi.org/10.1148/radiol.2223010766.
[4] M.B. Andersen, U. Bodtger, I.R. Andersen, et al., Metastases or benign adrenal lesions in patients with histopathological verification of lung cancer: can CT texture analysis distinguish? European Journal of Radiology 138 (2021) 109664 https://doi.org/10.1016/j.ejrad.2021.109664.
[5] A. Stanzione, R. Galatola, R. Cuocolo, et al., Radiomics in cross-sectional adrenal imaging: a systematic review and quality assessment study, Diagnostics 12 (3) (2022) 578, https://doi.org/10.3390/diagnostics12030578.
[6] Swetter Justin, M. Susan, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118, https://doi.org/10.1038/nature21056.
[7] H.D. Couture, L.A. Williams, G. Joseph, et al., Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype, NPJ Breast Cancer 4 (2018) 30, https://doi.org/10.1038/s41523-018-0079-1.
[8] B. Feng, X. Chen, Y. Chen, et al., Identifying solitary granulomatous nodules from solid lung adenocarcinoma: exploring robust image features with cross-domain transfer learning, Cancers (Basel) 15 (3) (2023) 892, https://doi.org/10.3390/cancers15030892.
[9] M. Kusunoki, T. Nakayama, A. Nishie, et al., A deep learning-based approach for the diagnosis of adrenal adenoma: a new trial using CT, BRIT J RADIOL 95 (1135) (2022) 20211066, https://doi.org/10.1259/bjr.20211066.
[10] Y. Singh, Z.S. Kelm, S. Faghani, et al., Deep learning approach for differentiating indeterminate adrenal masses using CT imaging, ABDOM RADIOL 48 (10) (2023) 3189–3194, https://doi.org/10.1007/s00261-023-03988-w.
[11] P. Alimu, C. Fang, Y. Han, et al., Artificial intelligence with a deep learning network for the quantification and distinction of functional adrenal tumors based on contrast-enhanced CT images, QUANT IMAG MED SURG 13 (4) (2023) 2675–2687, https://doi.org/10.21037/qims-22-539.
[12] C.C. Xiong, S.S. Zhu, D.H. Yan, et al., Rapid and precise detection of cancers via label-free SERS and deep learning, ANAL BIOANAL CHEM 415 (17) (2023) 3449–3462, https://doi.org/10.1007/s00216-023-04730-7.
[13] K. Yasaka, O. Abe, Deep learning and artificial intelligence in radiology: current applications and future directions, PLOS Med 15 (11) (2018) e1002707, https://doi.org/10.1371/journal.pmed.1002707.
[14] J. De Fauw, J.R. Ledsam, B. Romera-Paredes, et al., Clinically applicable deep learning for diagnosis and referral in retinal disease, Nat Med 24 (9) (2018) 1342–1350, https://doi.org/10.1038/s41591-018-0107-6.
[15] F. Fotouhi, A. Balu, Z. Jiang, et al., Dominating Set Model Aggregation for communication-efficient decentralized deep learning, NEURAL NETWORKS 171 (2023) 25–39, https://doi.org/10.1016/j.neunet.2023.11.057.

[16] E. Darzidehkalani, M. Ghasemi-Rad, P.M.A. van Ooijen, Federated learning in medical imaging: Part I: toward multicentral health care ecosystems, J AM COLL RADIOL 19 (8) (2022) 969–974, https://doi.org/10.1016/j.jacr.2022.03.015.
[17] Li Tian, Anit Kumar Sahu, Manzil Zaheer, et al., Federated optimization in heterogeneous networks, Machine Learning (2018), https://doi.org/10.48550/arXiv.1812.06127.
[18] X. Li, Z. Fang, Z. Shi, MpFedcon : model-contrastive personalized federated learning with the class center wuhan univ, J. Nat. Sci.. 27 (6) (2023) 508–520, https://doi.org/10.1051/wujns/2022276508.
[19] Meirui Jiang, Zirui Wang, Dou Qi, HarmoFL: harmonizing local and global drifts in federated learning on heterogeneous medical images, Image and Video Processing (2022), https://doi.org/10.48550/arXiv.2112.10775.
[20] Z. Yang, H. Wang, G. Kang, et al., Federated learning for breast density classification, IEEE Transactions on Medical Imaging 39 (5) (2020) 1555–1565.
[21] T. Li, A. Ghandeharioun, M.S. Rahman, et al., Privacy-preserving and robust deep learning for multisite and multivendor medical image classification, IEEE Transactions on Medical Imaging 39 (6) (2020) 1886–1898.
[22] Micah J. Sheller, G Anthony Reina, Brandon Edwards, et al., Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation, Brainlesion Workshop 1 (2) (2018) 92–104.
[23] L.M. Ho, E. Samei, M.A. Mazurowski, et al., Can texture analysis Be used to distinguish benign from malignant adrenal nodules on unenhanced CT, contrast-enhanced CT, or in-phase and opposed-phase MRI? AJR Am J Roentgenol 212 (3) (2019) 554–561, https://doi.org/10.2214/AJR.18.20097.
[24] M. Nakajo, M. Jinguji, M. Nakajo, et al., Texture analysis of FDG PET/CT for differentiating between FDG-avid benign and metastatic adrenal tumors: efficacy of combining SUV and texture parameters, Abdom Radiol (NY) 42 (12) (2017) 2882–2889, https://doi.org/10.1007/s00261-017-1207-3.
[25] B. Feng, X. Chen, Y. Chen, et al., Solitary solid pulmonary nodules: a CT-based deep learning nomogram helps differentiate tuberculosis granulomas from lung adenocarcinomas, Eur Radiol 30 (12) (2020) 6497–6507, https://doi.org/10.1007/s00330-020-07024-z.
[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al., Deep residual learning for image recognition, Computer Vision and Pattern Recognition (2015), https://doi.org/10.48550/arXiv.1512.03385.
[27] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, Computer Vision and Pattern Recognition (2014), https://doi.org/10.48550/arXiv.1409.1556.
[28] T. Li, A.K. Sahu, M. Zaheer, et al., Federated Optimization in Heterogeneous Networks 2 (2018) 429–450, https://doi.org/10.48550/arXiv.1812.06127.
[29] T. Dinh, Canh, Nguyen Tran, Josh Nguyen, Personalized federated learning with moreau envelopes, Advances in Neural Information Processing Systems 33 (2020) 21394–21405.
[30] S. Morita, Y. Nishina, H. Yamazaki, et al., Dual adrenal venous phase contrast-enhanced MDCT for visualization of right adrenal veins in patients with primary aldosteronism, EUR RADIOL 26 (7) (2015) 2073–2077, https://doi.org/10.1007/s00330-015-4073-9.
[31] K.R. Laukamp, R. Kessner, S. Halliburton, et al., Virtual noncontrast images from portal venous phase spectral-detector CT acquisitions for adrenal lesion characterization, J COMPUT ASSIST TOMO 45 (1) (2021) 24–28, https://doi.org/10.1097/RCT.0000000000000982.
[32] T.M. Kim, S.J. Choi, J.Y. Ko, et al., Fully automatic volume measurement of the adrenal gland on CT using deep learning to classify adrenal hyperplasia, Eur Radiol 33 (6) (2023) 4292–4302, https://doi.org/10.1007/s00330-022-09347-5.
[33] M. Hassaan, N. Ahmad, R.N. Ali, et al., DMFL_Net: a federated learning-based framework for the classification of COVID-19 from multiple chest diseases using X-rays, Sensors 23 (2) (2023) 743, https://doi.org/10.3390/s23020743.