Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Flexible and Scalable Federated Learning with Deep Feature Prompts for Digital Pathology

Cong Cong

thomas.cong@mq.edu.au

Macquarie University

Yang Song

UNSW Sydney

Antonio Di Ieva

Macquarie University

Angela Chou

NSW Health Pathology

Anthony J Gill

NSW Health Pathology

Enrico Coiera

Macquarie University

Sidong Liu

Macquarie University

Article

Keywords: WSI Classification, Federated Learning, Visual Prompt

Posted Date: August 8th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7060589/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at npj Digital Medicine on May 14th, 2026.
See the published version at https://doi.org/10.1038/s41746-026-02710-6.

Flexible and Scalable Federated Learning with Deep Feature Prompts for Digital Pathology

Cong Cong[1*], Yang Song[2], Antonio Di Ieva[3], Angela Chou[4,5], Anthony J Gill[4,5], Enrico Coiera[1], Sidong Liu[1]

[1*]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2113, Australia.
[2]School of Computer Science and Engineering, University of New South Wales, Kensington, NSW 2033, Australia.
[3]Computational NeuroSurgery (CNS) Lab, Macquarie Medical School, Macquarie University, Sydney, NSW 2113, Australia.
[4]Department of Anatomical Pathology, Royal North Shore Hospital, NSW Health Pathology, St Leonards, NSW 2065, Australia.
[5]Sydney Medical School, The University of Sydney, Sydney NSW 2006, Australia.

*Corresponding author(s). E-mail(s): thomas.cong@mq.edu.au;
Contributing authors: yang.song1@unsw.edu.au;
antonio.diieva@mq.edu.au; Angelashihyuan.chou@health.nsw.gov.au;
anthony.gill@health.nsw.gov.au; enrico.coiera@mq.edu.au;
sidong.liu@mq.edu.au;

Abstract

Collaborative learning across medical institutions is considered essential for developing robust and generalisable models in digital pathology. Federated learning (FL) offers a promising framework for enabling collaboration without the need to centralise data. Still, its practical adoption remains limited due to challenges such as high communication overhead, model heterogeneity across institutions, and privacy concerns. To address these challenges, we propose Federated Deep Feature Prompting (FedDFP), a novel and efficient FL paradigm designed to operate in real-world, heterogeneous clinical environments. FedDFP introduces a lightweight, learnable, client-specific prompt applied to patch-level embeddings extracted from whole slide images. By sharing only these compact prompts

across clients (*i.e.*, participating centres), FedDFP drastically reduces communication costs by over 99.9% compared to standard FL frameworks, with improved classification accuracy across clients. Through comprehensive evaluations on three public datasets (TCGA-IDH, CAMELYON16, and CAMELYON17), which include breast cancer and glioma classification tasks, FedDFP consistently outperforms both standard and personalised FL baselines. It maintains strong performance across diverse feature extractors and heterogeneous MIL classifiers, demonstrating flexibility, scalability, and robustness. These results highlight FedDFP as a promising solution for efficient, accurate, and privacy-conscious federated learning in digital pathology.

Keywords: WSI Classification, Federated Learning, Visual Prompt

1 Introduction

Digital pathology has emerged as an essential component in modern clinical workflows, offering standardised, efficient, and reproducible approaches to the microscopic examination of tissue specimens [1, 2]. The digitisation of pathology slides as Whole Slide Images (WSIs) enables computational algorithms, including computational fractal-based and deep learning-based methods, to assist pathologists in diagnostic tasks, improve diagnostic accuracy, reduce clinical workloads, and accelerate decision-making processes [3–7].

Despite these advances, the analysis of WSIs presents distinct computational challenges. WSIs are extremely high-resolution images, often comprising billions of pixels, and typically lack detailed region-level annotations, providing only slide-level diagnostic labels. Due to the impracticality of manually labelling every region within WSIs, computational pathology relies often on Multiple Instance Learning (MIL) frameworks. MIL methods treat each WSI as a collection of smaller image patches, which are individually processed into feature representations by a deep learning-based feature extractor. These extracted patch-level features are then aggregated through specialised MIL mechanisms to generate accurate slide-level predictions [6, 8–12].

Like many machine learning applications, the performance and robustness of MIL models in digital pathology depend critically on the availability of extensive, diverse, and high-quality training datasets [4]. Individual institutions often have limited datasets, particularly in the context of rare diseases or smaller hospitals in rural regions, thereby restricting the capacity of locally trained MIL models to generalise effectively on new cases. Collaborative learning across multiple institutions, where diverse and large-scale datasets are collectively leveraged, has become essential for improving the performance and generalisability of MIL models in clinical pathology applications [13]. However, sharing medical data poses challenges, including regulatory and legal constraints, as well as technical obstacles like the high costs of data transfer and storage [14, 15].

Federated Learning (FL) offers a promising solution for collaboratively collaborative training of machine learning models directly on distributed datasets, while

ensuring that data remain securely stored at individual participating centres (*e.g.*, hospitals, research institutions, or pathology labs). It does not require transferring sensitive patient data, thus respecting patient privacy and adhering to data-sharing regulations [16]. FL-based WSI classification methods have been actively developed in recent years [13, 17, 18]. For example, SiloBN [19] introduced local-statistic batch normalisation layers where the centre exchanges batch normalisation statistics with the server. Moreover, Zhang *et al.* [20] proposed generating pseudo histopathology patches locally and sharing them in the central server to train a more robust global model using self-supervised learning. Although these approaches have shown the effectiveness of adopting FL for WSI classification, they focus on patch-level classification, which requires precise patch-level annotations. However, such annotations are usually hard to obtain in real clinical scenarios. Thus, FedHisto [13] was proposed to tackle the more practical slide-level classification task. The MIL classifiers are shared between the centres and server using a FedAvg [21] framework. Huang *et al.* [17] further enhanced the local model performance by introducing a dynamic parameter updating mechanism.

Nevertheless, implementing these FLs in real clinical pathology scenarios involves substantial practical challenges: 1) *Heterogeneity Across Clients*: Different clients often vary considerably in computational resources, model preferences, and data characteristics [22]. Standard FL methods assume a uniform model architecture and identical computational resources across clients, thus restricting their applicability. Hospitals may prefer different neural network architectures, such as convolutional neural networks (CNNs) or Vision Transformers (ViTs), or might implement distinct MIL classifiers tailored to their specific diagnostic needs; 2) *Limited Communication Efficiency*: FL traditionally requires transferring large models, leading to high bandwidth usage [13, 17]. This is particularly burdensome for smaller, rural, or resource-constrained hospitals with limited network capabilities. The frequent exchange of full model parameters can be prohibitively expensive or slow, hindering the practical deployment of federated pathology solutions; 3) *Privacy Concerns and Regulatory Compliance*: Although federated learning significantly reduces privacy risks compared to direct data sharing, the repeated exchange of model parameters or gradients still poses potential privacy vulnerabilities. Recent studies have demonstrated that sensitive patient information can sometimes be inferred or reconstructed from model updates shared during FL [23, 24], thereby raising concerns about inadvertent privacy breaches.

To address the challenge of heterogeneity, personalised federated learning (pFL) methods have been proposed, allowing each client to train a customised model. For instance, decoupled training methods [25–27] divide models into shared layers for feature extraction and personalised layers for classification [25, 26] or use distinct models for local training and global aggregation [27]. More recent methods [28–32] exchange synthetic data with the central server rather than model parameters. However, these approaches typically demand substantial local computational resources and large volumes of data, making them less suitable for WSI classification. Another approach involves sharing class abstractions, such as local logits or prototypes [33–35]. While this approach alleviates high computational demands, different MIL classifiers may use distinct aggregation methods, resulting in inconsistent dimensions of class abstracts.

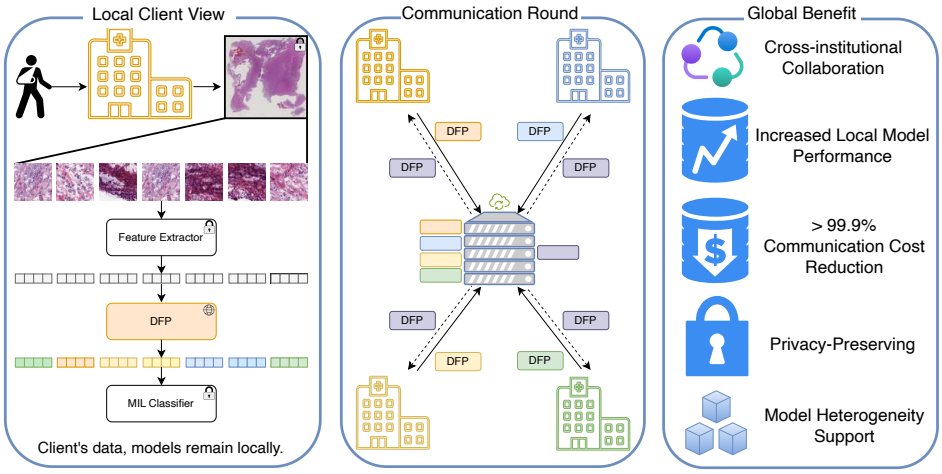


Fig. 1 Overall view of the proposed FedDFP. FedDFP is a communication-efficient and privacy-preserving federated learning framework, where each institution trains a local multiple instance learning (MIL) model using deep feature prompts (DFPs) that interact directly with patch-level features extracted from whole slide images (WSIs). Only the lightweight, client-specific DFPs are shared across institutions for aggregation, enabling collaborative learning without exchanging raw data or model parameters.

In contrast, recent pFL methods have incorporated visual prompts (VP) [15, 36–38]; locally tuned VPs are treated as local abstracts and shared with the central server. Since VPs operate in the input space, their dimensionality remains fixed and independent of the underlying classifier architecture. This allows different clients to collaborate effectively despite using heterogeneous MIL models, as the prompt vectors can be aggregated without requiring alignment of internal feature or output representations. However, their effectiveness in WSI classification remains unexplored. Therefore, this work seeks to develop a VP-based pFL framework for WSI classification to address this gap.

To simultaneously overcome the challenges of client heterogeneity, communication efficiency, and privacy in multi-institutional collaborative learning, this work proposes the Federated Deep Feature Prompt (FedDFP), a novel federated learning framework leveraging the client-specific deep feature prompts (DFP) as shown in Fig. 1. Specifically, WSIs at each centre are converted into patch-level feature embeddings using a pre-trained feature extractor of their choice (*e.g.*, CNN [39] or ViT [40]). A lightweight, learnable, client-specific prompt vector is then applied directly to these patch-level feature embeddings, effectively refining them for better task-specific representation. These prompt-refined features are subsequently aggregated by an MIL model to produce slide-level predictions. FedDFP facilitates communication-efficient collaborative training by exchanging only these compact feature prompts, significantly reducing transmission overhead. Furthermore, our framework is designed with privacy considerations in mind, as it facilitates the exchange of only lightweight visual prompts between institutions, rather than sensitive patient data or full model parameters. Extensive experiments on diverse datasets, including TCGA-IDH for glioma classification and

CAMELYON16/17 for metastatic breast cancer detection, show that FedDFP consistently achieves state-of-the-art performance, improves per-centre diagnostic accuracy, and offers seamless integration into varied computational workflows without sacrificing communication or privacy efficiency.

2 Results

This section is structured as follows: first, we demonstrate the effectiveness of federated learning in improving the performance of local models. We then evaluate the performance of our proposed FedDFP framework, specifically addressing two critical real-world clinical challenges: communication efficiency and inter-client heterogeneity. All analyses and validations were conducted using three publicly available datasets covering two important clinical tasks: breast cancer metastasis detection and glioma IDH genetic subtype classification.

2.1 Federated Learning Improves Model Performance Across Institutions

To evaluate the effectiveness of federated learning (FL) in real-world digital pathology scenarios, we compared a broad set of FL methods against isolated (local) model training across the benchmark datasets. Two major groups of FL baselines were considered. The first group consists of standard FL methods, assuming a globally shared model is sufficient for all participant ce ntres (*e.g.*, FedHisto [13], an enhanced version of FedAvg [21] tailored to WSI classification, FedDyn [41], MOON [42] and SCAFFOLD [43]). The second group comprises personalised FL methods, which explicitly account for inter-client heterogeneity and aim to tailor models to individual clients (*e.g.*, FedProto [35], DESA [28], SGPT [31] and pFedPG [37]).

We first present the per-centre classification accuracy analysis, which shows that federated learning improves classification performance across nearly all participating centres, as demonstrated in Table 1 and Table 2, using ResNet and ViT as feature extractors, respectively. For example, in the IDH dataset using the ResNet50 feature extractor (Table 1), local models performed poorly for centres with limited data (*e.g.*, C7: 50.0%), but FL methods substantially improved performance (FedDyn: 86.7%, FedProto: 81.9%, DESA: 83.3%). Similar trends were observed across CA17, where several centres (*e.g.*, C2, C4) benefited from federated training with gains of over 10%.

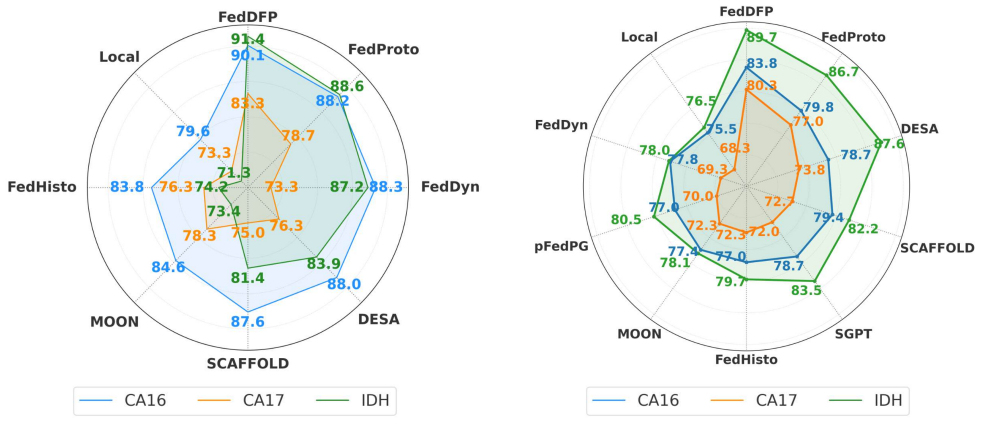
We also compared the averaged classification accuracy over all centres in Fig. 2 (a) and (b), both standard and personalised FL approaches consistently outperformed local training across datasets and model backbones. On CA16 (ResNet50), the average local accuracy was 73.3%, compared to 76.3% (FedHisto), 78.3% (MOON), and 78.7% (FedProto). For the more heterogeneous IDH dataset, local models averaged 71.3%, while FL methods achieved 74.2% (FedHisto), 87.2% (FedDyn), 83.9% (DESA), and 88.6% (FedProto). These improvements were mirrored with ViT backbones, underscoring the consistency of FL benefits across feature extractors. Additionally, the ROC curves and AUC scores in Fig. 3 (a-c) indicate that FL methods achieve higher AUC scores than pure local training, indicating better class discrimination. The ROC

Methods		Local	FedHisto	FedDyn	MOON	SCAFFOLD	FedProto	DESA	FedDFP (Ours)
			Standard FL				*Personalised FL*		
CA16	C1	$85.2_{\pm 1.0}$	$\mathbf{92.8}_{\pm 1.7}$	$92.3_{\pm 0.6}$	$91.0_{\pm 1.7}$	$89.6_{\pm 1.7}$	$92.8_{\pm 0.6}$	$92.3_{\pm 0.6}$	$92.3_{\pm 0.6}$
	C2	$74.0_{\pm 1.6}$	$74.8_{\pm 1.8}$	$84.3_{\pm 1.8}$	$78.0_{\pm 1.7}$	$85.5_{\pm 0.9}$	$83.6_{\pm 0.9}$	$83.6_{\pm 0.9}$	$\mathbf{88.1}_{\pm 0.9}$
	p-value	1.15e-06	2.32e-04	1.00e-03	1.46e-06	7.39e-04	2.54e-03	1.35e-03	-
CA17	C1	$80.0_{\pm 0.0}$	$85.0_{\pm 4.1}$	$80.0_{\pm 0.0}$	$\mathbf{90.0}_{\pm 0.0}$	$83.3_{\pm 2.4}$	$88.3_{\pm 2.4}$	$90.0_{\pm 0.0}$	$90.0_{\pm 0.0}$
	C2	$60.0_{\pm 4.1}$	$71.7_{\pm 2.4}$	$71.7_{\pm 2.4}$	$61.7_{\pm 2.4}$	$61.7_{\pm 2.4}$	$75.0_{\pm 4.1}$	$75.0_{\pm 0.0}$	$\mathbf{75.0}_{\pm 7.1}$
	C3	$86.7_{\pm 2.4}$	$81.7_{\pm 2.4}$	$83.3_{\pm 4.7}$	$83.3_{\pm 2.4}$	$86.7_{\pm 2.4}$	$85.0_{\pm 0.0}$	$76.6_{\pm 4.7}$	$\mathbf{88.3}_{\pm 2.4}$
	C4	$60.0_{\pm 0.0}$	$61.7_{\pm 2.4}$	$58.3_{\pm 2.4}$	$63.3_{\pm 2.4}$	$58.7_{\pm 2.4}$	$65.0_{\pm 0.0}$	$61.7_{\pm 2.4}$	$\mathbf{81.7}_{\pm 2.4}$
	C5	$80.0_{\pm 0.0}$	$81.7_{\pm 4.7}$	$78.3_{\pm 2.4}$	$81.7_{\pm 2.4}$	$80.0_{\pm 0.0}$	$80.0_{\pm 0.0}$	$78.3_{\pm 2.4}$	$\mathbf{81.7}_{\pm 2.4}$
	p-value	8.43e-04	1.97e-03	6.61e-04	2.83e-03	3.79e-04	4.74e-03	2.58e-03	-
IDH	C1	$89.5_{\pm 2.2}$	$89.8_{\pm 2.1}$	$84.0_{\pm 2.4}$	$67.9_{\pm 1.6}$	$\mathbf{94.5}_{\pm 2.6}$	$83.5_{\pm 2.1}$	$79.8_{\pm 4.7}$	$94.5_{\pm 1.6}$
	C2	$64.6_{\pm 3.0}$	$66.7_{\pm 2.9}$	$87.5_{\pm 5.1}$	$66.7_{\pm 5.6}$	$77.1_{\pm 5.6}$	$91.7_{\pm 7.8}$	$91.7_{\pm 3.0}$	$\mathbf{93.8}_{\pm 5.1}$
	C3	$80.0_{\pm 1.6}$	$70.0_{\pm 1.4}$	$86.7_{\pm 9.4}$	$76.7_{\pm 2.5}$	$80.0_{\pm 0.0}$	$\mathbf{96.7}_{\pm 4.7}$	$90.0_{\pm 0.0}$	$86.7_{\pm 4.7}$
	C4	$68.3_{\pm 2.4}$	$78.3_{\pm 6.2}$	$80.0_{\pm 7.1}$	$83.3_{\pm 4.7}$	$85.0_{\pm 4.1}$	$\mathbf{95.0}_{\pm 4.1}$	$81.7_{\pm 4.7}$	$91.7_{\pm 2.4}$
	C5	$66.7_{\pm 1.7}$	$86.1_{\pm 1.4}$	$90.3_{\pm 1.1}$	$86.7_{\pm 4.7}$	$90.3_{\pm 5.2}$	$90.0_{\pm 0.0}$	$68.1_{\pm 8.0}$	$\mathbf{91.7}_{\pm 3.4}$
	C6	$74.5_{\pm 7.8}$	$72.6_{\pm 3.7}$	$89.2_{\pm 3.7}$	$80.4_{\pm 5.5}$	$74.5_{\pm 3.7}$	$81.4_{\pm 7.0}$	$83.3_{\pm 5.6}$	$\mathbf{91.2}_{\pm 2.4}$
	C7	$50.0_{\pm 1.6}$	$50.0_{\pm 1.6}$	$86.7_{\pm 4.7}$	$45.8_{\pm 6.8}$	$66.7_{\pm 9.4}$	$81.9_{\pm 2.0}$	$83.3_{\pm 4.7}$	$\mathbf{90.0}_{\pm 8.2}$
	C8	$71.7_{\pm 9.4}$	$80.0_{\pm 1.5}$	$93.3_{\pm 2.4}$	$85.0_{\pm 0.0}$	$83.3_{\pm 9.4}$	$88.3_{\pm 9.4}$	$93.3_{\pm 2.4}$	$\mathbf{96.7}_{\pm 2.4}$
	p-value	6.67e-08	1.03e-04	5.59e-04	2.81e-06	3.38e-04	1.62e-03	5.05e-05	-
#Com		-	791K	791K	791K	791K	0.5K	24M	**1.02K**

Table 1 Performance comparison of FL methods on CAMELYON16 (**CA16**), CAMELYON17 (**CA17**), and TCGA-IDH (**IDH**), using *ResNet50* as the patch-level feature extractor and CLAM [10] as the MIL classifier. The table reports centre-wise accuracy and corresponding p-values. Communication cost (**#Com**), measured by the number of parameters exchanged per round, is also listed for each method. Best-performing results are highlighted in **bold**.

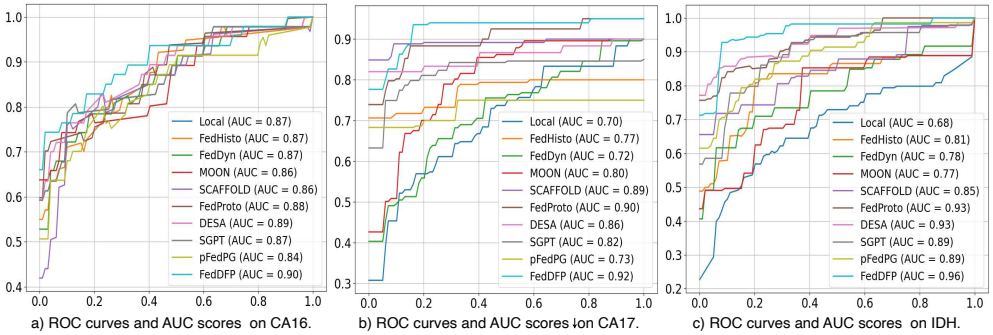
Methods		Local	FedHisto	FedDyn	MOON	SCAFFOLD	FedProto	DESA	SGPT	pFedPG	FedDFP (Ours)
				Standard FL					*Personalised FL*		
CA16	C1	$73.6_{\pm 1.5}$	$78.3_{\pm 1.9}$	$85.6_{\pm 2.3}$	$80.0_{\pm 0.7}$	$82.7_{\pm 2.6}$	$84.4_{\pm 0.7}$	$84.2_{\pm 1.7}$	$80.1_{\pm 0.3}$	$79.3_{\pm 0.6}$	$\mathbf{86.0}_{\pm 1.2}$
	C2	$77.5_{\pm 5.2}$	$75.0_{\pm 2.1}$	$70.0_{\pm 2.6}$	$74.8_{\pm 1.8}$	$76.1_{\pm 1.8}$	$75.4_{\pm 0.0}$	$73.1_{\pm 0.5}$	$77.4_{\pm 0.0}$	$75.0_{\pm 1.8}$	$\mathbf{80.5}_{\pm 2.2}$
	p-value	3.73e-03	8.43e-04	1.55e-03	6.47e-04	6.14e-04	5.05e-04	1.30e-03	1.22e-04	9.92e-05	-
CA17	C1	$56.7_{\pm 4.7}$	$81.3_{\pm 2.4}$	$56.7_{\pm 4.7}$	$81.7_{\pm 8.5}$	$78.3_{\pm 2.4}$	$80.0_{\pm 4.1}$	$66.7_{\pm 2.4}$	$71.7_{\pm 4.7}$	$65.0_{\pm 4.1}$	$\mathbf{80.0}_{\pm 4.1}$
	C2	$63.3_{\pm 2.4}$	$66.7_{\pm 2.4}$	$78.3_{\pm 2.4}$	$65.0_{\pm 0.0}$	$68.3_{\pm 2.4}$	$68.3_{\pm 2.4}$	$70.0_{\pm 4.1}$	$66.7_{\pm 2.4}$	$61.7_{\pm 6.2}$	$\mathbf{75.0}_{\pm 4.1}$
	C3	$83.3_{\pm 2.4}$	$86.7_{\pm 2.4}$	$85.0_{\pm 4.1}$	$86.7_{\pm 2.4}$	$83.3_{\pm 2.4}$	$88.3_{\pm 2.4}$	$85.0_{\pm 0.0}$	$90.0_{\pm 4.1}$	$81.7_{\pm 2.4}$	$\mathbf{91.7}_{\pm 2.4}$
	C4	$76.7_{\pm 2.4}$	$63.3_{\pm 4.7}$	$71.7_{\pm 6.2}$	$66.7_{\pm 2.4}$	$68.3_{\pm 6.2}$	$71.7_{\pm 6.2}$	$78.3_{\pm 4.1}$	$60.0_{\pm 0.0}$	$76.7_{\pm 6.2}$	$\mathbf{80.0}_{\pm 0.0}$
	C5	$76.7_{\pm 2.4}$	$61.7_{\pm 6.2}$	$65.0_{\pm 8.2}$	$61.7_{\pm 2.4}$	$65.0_{\pm 4.1}$	$75.0_{\pm 4.1}$	$68.3_{\pm 4.1}$	$70.0_{\pm 4.1}$	$65.0_{\pm 7.1}$	$\mathbf{80.0}_{\pm 0.0}$
	p-value	8.07e-06	1.68e-04	9.06e-06	2.56e-04	2.91e-03	1.17e-02	2.57e-07	1.92e-04	8.44e-06	-
IDH	C1	$69.6_{\pm 1.8}$	$74.3_{\pm 6.2}$	$85.2_{\pm 6.6}$	$94.5_{\pm 1.6}$	$94.1_{\pm 3.0}$	$84.0_{\pm 4.3}$	$82.7_{\pm 5.3}$	$79.8_{\pm 4.5}$	$75.1_{\pm 1.6}$	$\mathbf{83.5}_{\pm 4.5}$
	C2	$69.0_{\pm 8.4}$	$85.4_{\pm 3.0}$	$79.2_{\pm 7.8}$	$72.9_{\pm 7.8}$	$85.4_{\pm 5.9}$	$93.8_{\pm 5.1}$	$91.7_{\pm 7.8}$	$85.4_{\pm 9.6}$	$95.8_{\pm 5.9}$	$\mathbf{97.9}_{\pm 2.1}$
	C3	$86.7_{\pm 4.7}$	$83.3_{\pm 4.7}$	$80.0_{\pm 8.2}$	$63.3_{\pm 8.6}$	$76.7_{\pm 4.7}$	$90.0_{\pm 8.2}$	$90.0_{\pm 8.2}$	$85.0_{\pm 8.2}$	$86.7_{\pm 4.7}$	$\mathbf{90.0}_{\pm 2.8}$
	C4	$68.3_{\pm 6.2}$	$83.3_{\pm 4.7}$	$81.7_{\pm 8.5}$	$81.7_{\pm 4.7}$	$90.0_{\pm 4.1}$	$85.0_{\pm 7.1}$	$85.0_{\pm 0.0}$	$78.3_{\pm 9.3}$	$83.3_{\pm 4.7}$	$\mathbf{93.3}_{\pm 6.2}$
	C5	$80.4_{\pm 1.4}$	$90.0_{\pm 8.2}$	$77.8_{\pm 3.9}$	$69.4_{\pm 5.2}$	$84.7_{\pm 4.0}$	$72.2_{\pm 3.9}$	$72.2_{\pm 8.6}$	$65.8_{\pm 5.2}$	$58.3_{\pm 5.9}$	$\mathbf{80.6}_{\pm 5.2}$
	C6	$51.4_{\pm 2.0}$	$81.4_{\pm 9.7}$	$66.7_{\pm 9.1}$	$73.5_{\pm 4.2}$	$76.5_{\pm 8.7}$	$88.2_{\pm 4.2}$	$87.3_{\pm 6.0}$	$84.3_{\pm 2.8}$	$76.5_{\pm 4.8}$	$\mathbf{88.2}_{\pm 2.4}$
	C7	$96.7_{\pm 4.7}$	$51.4_{\pm 7.1}$	$56.7_{\pm 4.7}$	$80.0_{\pm 8.2}$	$66.7_{\pm 4.7}$	$90.0_{\pm 4.4}$	$96.7_{\pm 4.7}$	$96.6_{\pm 4.7}$	$86.7_{\pm 4.7}$	$\mathbf{96.7}_{\pm 2.7}$
	C8	$86.7_{\pm 4.7}$	$83.3_{\pm 4.7}$	$70.0_{\pm 10.8}$	$66.8_{\pm 2.5}$	$83.3_{\pm 6.2}$	$85.0_{\pm 0.0}$	$85.0_{\pm 7.1}$	$84.3_{\pm 2.8}$	$81.7_{\pm 6.2}$	$\mathbf{88.3}_{\pm 8.5}$
	p-value	3.94e-04	9.14e-05	2.06e-04	1.14e-05	3.19e-04	2.92e-02	6.36e-06	4.19e-05	6.84e-07	-
#Com		-	125K	125K	125K	125K	0.2K	24M	0.4K	0.4K	**0.4K**

Table 2 Performance comparison of FL methods on CAMELYON16 (**CA16**), CAMELYON17 (**CA17**), and TCGA-IDH (**IDH**), using *ViT_S_16* as the patch-level feature extractor and CLAM [10] as the MIL classifier. The table reports centre-wise accuracy and corresponding p-values. Communication cost (**#Com**), measured by the number of parameters exchanged per round, is also listed for each method. Best-performing results are highlighted in **bold**.



a) Accuracy Comparison using ResNet as Feature Extractor. b) Accuracy Comparison using ViT as Feature Extractor.

Fig. 2 Averaged slide-level classification accuracy comparison across three benchmark datasets, CAMELYON16 (CA16), CAMELYON17 (CA17), and TCGA-IDH (IDH), using 7 (ResNet as Feature Extractor) or 9 (ViT as Feature Extractor) baseline methods and our proposed FedDFP.



a) ROC curves and AUC scores on CA16. b) ROC curves and AUC scores on CA17. c) ROC curves and AUC scores on IDH.

Fig. 3 The ROC curves and AUC scores comparison across three benchmark datasets, CAMELYON16 (CA16), CAMELYON17 (CA17), and TCGA-IDH (IDH), using 7 (ResNet as Feature Extractor) or 9 (ViT as Feature Extractor) baseline methods and our proposed FedDFP.

curve further supports this, remaining consistently close to the top-left corner across thresholds, indicating robust performance at varied decision boundaries.

2.2 FedDFP Outperforms Other FL Methods

As shown in Fig. 2 (a) and (b), our proposed FedDFP framework consistently achieved superior performance across all datasets and feature extractors. FedDFP outperformed both standard FL approaches (such as FedHisto, FedDyn, MOON, and SCAFFOLD) and recent personalised methods (such as FedProto, DESA, SGPT, and pFedPG), demonstrating its ability to adapt to inter-institutional variability while maintaining strong generalisation. On the CA16 dataset, FedDFP achieved the highest average accuracy of 90.1% (ResNet50) and 83.8% (ViT), surpassing the next best performing methods, FedDyn 88.3% (ResNet50) and FedProto 79.8% (ViT). Similarly, on CA17,

FedDFP led with 83.3% (ResNet50) and 80.3% (ViT), significantly outperforming FedProto and SGPT by 2–5%. The gains were most notable on the more challenging IDH dataset, where FedDFP reached 91.4% (ResNet50) and 89.7% (ViT) in average accuracy, outperforming the best baseline (FedProto, 88.6%) by a margin of nearly 3%. Importantly, per-centre comparisons in Table 1 and Table 2 further highlight FedDFP's robustness. To verify the statistical significance of these improvements, we performed paired t-tests over five independent runs, with the null hypothesis assuming no significant difference between FedDFP and each baseline. Across all datasets and model backbones, the p-values were consistently below 0.05, confirming that the performance gains of FedDFP are statistically significant.

2.3 FedDFP Enables Highly Efficient Communication

While standard FL methods improve accuracy, they often require transmitting large model updates (*e.g.*, tens or hundreds of thousands of parameters) per communication round. This presents a major limitation in resource-constrained or bandwidth-limited clinical settings. FedDFP addresses this challenge by exchanging only a lightweight prompt vector, significantly reducing communication overhead. As shown in the last row of Table 1, FedDFP reduced communication costs from approximately 791K parameters (used by FedHisto [13] and SCAFFOLD [43]) to just 0.4K, representing a reduction of over 99.9%. Despite this drastic decrease, FedDFP still outperformed all baselines. This demonstrates that effective collaboration in federated learning can be achieved with minimal communication, making FedDFP particularly suitable for real-world deployments where bandwidth is a limiting factor.

Beyond lowering the per-round communication cost, FedDFP also accelerates convergence, requiring fewer communication rounds to reach optimal performance. As illustrated in Fig. 4, FedDFP converged much faster than other methods, typically reaching peak performance within 20 rounds, compared to 80–100 rounds for standard FL baselines. This further compounds its communication efficiency, reducing the total communication budget by one order of magnitude.

Another key factor in FedDFP's efficiency is the small number of prompts required to achieve strong performance. As demonstrated in Fig. 4, while the best-performing configuration may vary depending on the task, we find that using as few as a single prompt per client already provides significant gains over existing FL methods.

2.4 FedDFP Performs Robustly Across Feature Extractors

In multiple instance learning (MIL) pipelines for whole slide image (WSI) classification, the feature extractor plays a critical role in converting raw image patches into compact, high-level embeddings that are subsequently aggregated by the MIL classifier. The choice of feature extractor, whether convolutional (*e.g.*, ResNet [39]) or transformer-based (*e.g.*, ViT [40]), can substantially influence both the quality of extracted representations and the downstream model performance. In real-world clinical settings, different institutions may prefer different architectures based on their computational infrastructure, prior workflows, or existing deployment constraints. Therefore, ensuring that a federated learning method remains effective across diverse

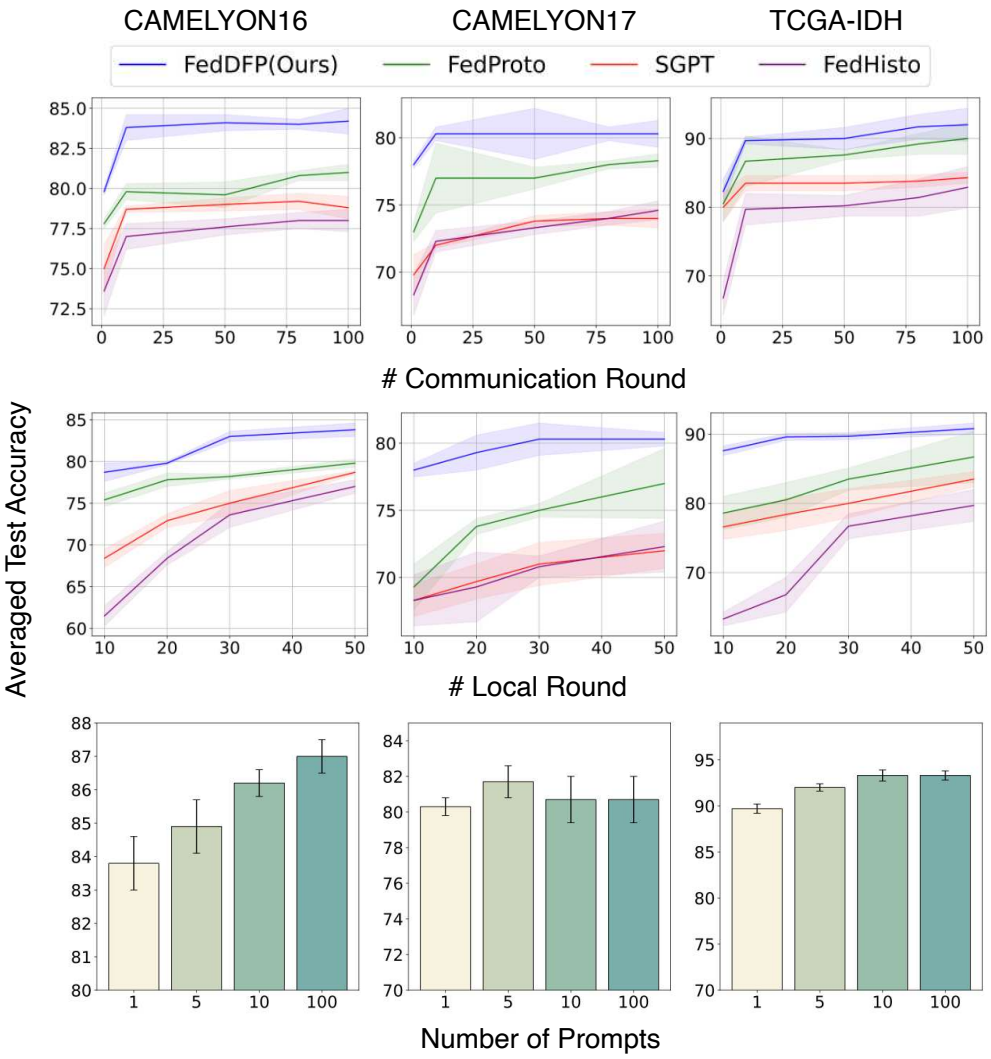


Fig. 4 Global averaged accuracy vs. different number of global communication rounds (**row 1**), different number of local training rounds (**row 2**) and number of prompts (**row 3**).

feature extractors is essential for practical adoption. To simulate this real-world heterogeneity, we conducted experiments using two distinct feature extractors: ResNet50 and ViT-S/16. For each run, we extracted patch-level features using one backbone and trained all MIL classifiers using only features derived from that extractor, ensuring consistency within each experiment. As shown in Table 1 and 2, across all datasets, FedDFP achieved consistent improvements over both local training and standard FL baselines, regardless of the backbone used. On CAMELYON16, FedDFP improved average accuracy from $79.6\% \pm 0.6$ (ResNet50) to $90.1\% \pm 0.2$, and from $75.5\% \pm 3.3$ (local ViT) to $83.8\% \pm 0.8$. A similar pattern held for TCGA-IDH, where FedDFP

	Methods		DESA	FedHE	SGPT	pFedPG	**FedDFP**
CA16	C1	TransMIL	$80.8_{\pm 2.6}$	$75.2_{\pm 2.2}$	$80.6_{\pm 1.2}$	$75.2_{\pm 4.2}$	$\mathbf{82.0}_{\pm 7.8}$
	C2	ABMIL	$67.6_{\pm 3.4}$	$66.1_{\pm 3.9}$	$71.1_{\pm 1.8}$	$75.5_{\pm 7.1}$	$\mathbf{75.5}_{\pm 4.1}$
	p-value	-	1.58e-03	2.36e-04	2.45e-02	6.01e-03	-
CA17	C1	TransMIL	$71.7_{\pm 6.2}$	$75.0_{\pm 0.0}$	$75.0_{\pm 0.0}$	$73.3_{\pm 6.2}$	$\mathbf{78.3}_{\pm 4.7}$
	C2	ABMIL	$70.0_{\pm 4.1}$	$61.7_{\pm 2.3}$	$68.3_{\pm 2.4}$	$66.7_{\pm 6.2}$	$\mathbf{68.3}_{\pm 2.4}$
	C3	CLAM	$93.3_{\pm 2.4}$	$80.0_{\pm 0.0}$	$88.3_{\pm 2.4}$	$88.3_{\pm 2.4}$	$\mathbf{95.0}_{\pm 0.0}$
	C4	TransMIL	$68.3_{\pm 2.4}$	$73.3_{\pm 2.4}$	$71.6_{\pm 6.2}$	$75.0_{\pm 4.1}$	$\mathbf{75.0}_{\pm 0.0}$
	C5	ABMIL	$75.0_{\pm 4.1}$	$75.0_{\pm 4.1}$	$75.0_{\pm 4.1}$	$71.7_{\pm 2.4}$	$\mathbf{80.0}_{\pm 4.1}$
	p-value	-	1.49e-03	7.48e-04	5.26e-04	1.07e-02	-
IDH	C1	TransMIL	$84.4_{\pm 5.7}$	$77.6_{\pm 1.6}$	$75.8_{\pm 2.2}$	$70.0_{\pm 5.7}$	$\mathbf{86.9}_{\pm 2.1}$
	C2	ABMIL	$89.6_{\pm 10.6}$	$93.8_{\pm 5.1}$	$83.3_{\pm 3.0}$	$83.3_{\pm 3.0}$	$\mathbf{93.8}_{\pm 4.7}$
	C3	CLAM	$90.0_{\pm 0.4}$	$90.0_{\pm 0.8}$	$90.0_{\pm 0.0}$	$90.0_{\pm 8.2}$	$\mathbf{90.0}_{\pm 0.0}$
	C4	TransMIL	$86.7_{\pm 4.8}$	$76.7_{\pm 6.2}$	$76.7_{\pm 4.7}$	$80.0_{\pm 7.1}$	$\mathbf{86.7}_{\pm 2.4}$
	C5	ABMIL	$72.2_{\pm 10.9}$	$66.7_{\pm 3.6}$	$76.4_{\pm 5.2}$	$62.5_{\pm 3.4}$	$\mathbf{91.7}_{\pm 5.9}$
	C6	CLAM	$87.3_{\pm 5.6}$	$84.3_{\pm 7.3}$	$80.4_{\pm 3.6}$	$85.3_{\pm 8.7}$	$\mathbf{87.3}_{\pm 5.0}$
	C7	TransMIL	$83.3_{\pm 4.7}$	$96.7_{\pm 3.2}$	$86.7_{\pm 4.7}$	$93.3_{\pm 4.7}$	$\mathbf{96.7}_{\pm 3.2}$
	C8	ABMIL	$91.7_{\pm 6.3}$	$91.7_{\pm 4.7}$	$83.3_{\pm 8.5}$	$93.3_{\pm 6.2}$	$\mathbf{93.3}_{\pm 4.7}$
	p-value	-	6.67e-03	3.57e-03	1.29e-04	4.04e-05	-

Table 3 Heterogeneous model performance comparison between FedDFP and *personalised FL* methods on CAMELYON16 (**CA16**), CAMELYON17 (**CA17**), and TCGA-IDH (**IDH**). We report the MIL method selected by each centre and the per-centre accuracy. The best results are highlighted in **bold**.

increased accuracy from $71.3\% \pm 2.5$ (ResNet50) to $91.4\% \pm 2.4$, and from $76.5\% \pm 3.6$ (ViT) to $89.7\% \pm 0.5$. Notably, the standard deviation in FedDFP results was consistently lower than in other methods, indicating improved training stability. These findings confirm that FedDFP generalises well across network architectures and clinical tasks, making it a flexible and robust solution for diverse digital pathology applications.

2.5 Addressing Client Setting Heterogeneity using FedDFP

In real-world clinical environments, it is common for different medical institutions to adopt varying machine learning models based on their computational resources, domain expertise, and deployment pipelines. This heterogeneity is particularly pronounced in multiple instance learning (MIL) based whole slide image (WSI) classification workflows, where each centre may prefer a different MIL architecture. To simulate this realistic scenario, we define a heterogeneous FL setup in which each medical centre independently adopts a classifier randomly selected from a pool of widely used MIL methods: CLAM [10], TransMIL [44], and ABMIL [45]. The details of MIL selection of each centre are shown in the third column in Table 3. We maintain a consistent ViT-S/16 backbone across all centres to ensure a common feature space and allow direct comparison to prior visual prompt-based methods. We compare FedDFP to a selection of personalised FL baselines that are compatible with heterogeneous models, including DESA [28], a synthetic data-sharing method; and two visual prompt-based techniques, including SGPT [31] and pFedPG [37]. Moreover, since class embedding dimensions vary across different MIL classifiers, FedProto cannot be applied. Instead, we use FedHE [34], which shares class logits, for comparison. These methods represent

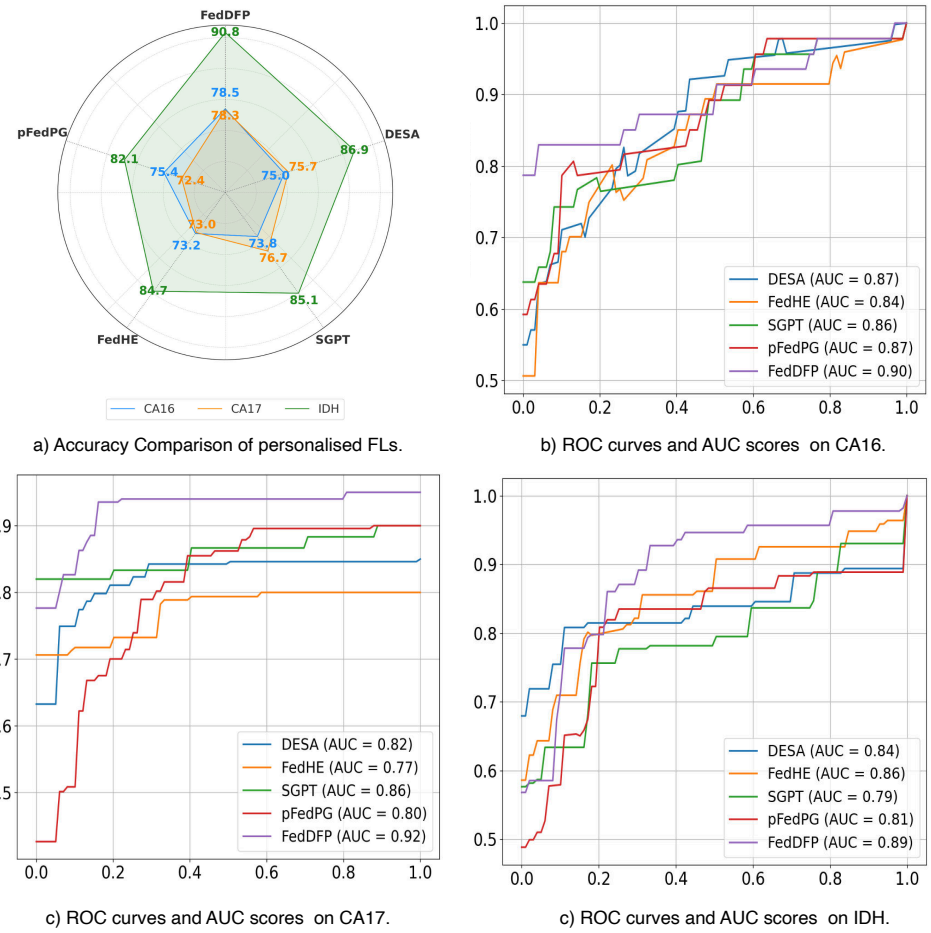


Fig. 5 Performance comparison across three benchmark datasets, CAMELYON16 (CA16), CAMELYON17 (CA17), and TCGA-IDH (IDH), between FedDFP and 4 personalised FLs. (a) The radar plot showing the average slide-level classification accuracy achieved by each method; (b-d) The ROC curves and AUC scores achieved by each method across the three datasets.

the state-of-the-art in enabling personalised collaboration across non-identical model architectures. As shown in Table 3 and Fig. 5, FedDFP consistently outperforms all baselines in both average classification accuracy and per-centre accuracy across three benchmark datasets. Specifically, it achieves average accuracies of 78.7% on CAMELYON16, 79.3% on CAMELYON17, and 90.8% on TCGA-IDH which outperforms all personalised FL methods in each case. These results highlight the effectiveness of FedDFP in real-world heterogeneous settings. By applying prompts at the feature embedding level, FedDFP enables architecture-agnostic collaboration: each centre can retain its preferred MIL design while still benefiting from shared knowledge. This flexibility, combined with superior accuracy, positions FedDFP as a highly practical and

scalable solution for federated digital pathology, where institutional variability is the norm rather than the exception.

3 Discussion

Collaborative learning across institutions is increasingly recognised as essential for developing robust and generalisable clinical AI models, especially in domains like digital pathology, where individual institutions may have limited or biased datasets. Compared to isolated local training, federated learning (FL) enables institutions to collaboratively train shared models without directly exchanging patient data, offering a powerful framework for privacy-preserving multi-centre AI development. However, existing FL methods often struggle to balance performance, efficiency, and adaptability in real-world settings, particularly when inter-client heterogeneity is pronounced. In digital pathology, this challenge is exacerbated by significant variations in data distribution across institutions, including differences in staining protocols, scanner types, and diagnostic practices. Traditional FL approaches, which typically rely on uniform model architectures and weight aggregation, are ill-suited to cope with such heterogeneity and often lead to suboptimal generalisation and client-level performance. Furthermore, large model sizes and frequent communication requirements impose practical burdens on clinical collaborators with limited computational resources.

In this context, we propose FedDFP, a novel paradigm for collaborative learning across heterogeneous clinical environments. By integrating lightweight feature-level prompts into the FL process, FedDFP significantly improves classification performance while drastically reducing communication costs. Our results demonstrate that FedDFP not only outperforms local models and standard FL baselines but also consistently exceeds the performance of advanced personalised FL methods. Notably, FedDFP maintains its advantages across different datasets, feature extractors, and MIL architectures, highlighting its robustness and practical suitability for deployment in diverse healthcare settings. To the best of our knowledge, this is the first work to introduce domain-specific feature prompting into federated learning for digital pathology, enabling each institution to condition the shared feature extractor using a personalised prompt without altering the local classifier or compromising privacy. FedDFP provides a flexible and modular framework that accommodates institutional heterogeneity which is a common barrier in clinical AI deployments. Furthermore, this work significantly advances the field of federated digital pathology by addressing critical trade-offs between generalisation, communication efficiency, and model adaptability. Unlike prior approaches that require full model sharing or computationally expensive personalization layers, FedDFP achieves strong performance using minimal parameter overhead, making it particularly well-suited for resource-constrained settings. As a result, this work may pave the way for practical, scalable, and privacy-preserving AI deployment in real-world multi-centre studies, offering a promising direction for future research in federated learning and clinical model translation. The modularity and efficiency of FedDFP suggest its potential as a generalisable strategy beyond pathology, applicable to other data-rich but privacy-sensitive domains such as radiology, genomics, and multi-omics integration.

A key finding from our evaluation is that FedDFP achieves these gains with dramatically reduced communication costs. By transmitting only a compact prompt vector, FedDFP reduces the per-round communication size by over 99.9% compared to conventional approaches, such as FedAvg-based methods or full model fine-tuning. Importantly, our ablation studies further show that FedDFP converges substantially faster, often reaching peak performance within 20 communication rounds, compared to 80–100 rounds required by baseline methods. Together, these improvements translate into a significantly lower total communication budget, which is especially important in bandwidth-constrained or resource-limited clinical environments, such as rural hospitals or cross-continental research collaborations, where internet infrastructure and hardware capabilities may be limited. This aspect aligns well with recent initiatives to bring AI into real-time and edge-assisted healthcare settings, where models must be both compact and quickly trainable without burdening the clinical workflow [46, 47]. For example, in large-scale consortia such as The Cancer Genome Atlas (TCGA) or The Australian Digital Health Agency, scalability and interoperability remain key technical barriers to integrating AI systems across geographically distributed institutions. FedDFP's lightweight design and fast convergence are particularly suitable for these scenarios, where continual model updates and secure communications must occur in parallel with time-sensitive clinical care.

In addition to performance and efficiency, FedDFP offers several privacy-preserving advantages. Since only feature-level prompts are exchanged, FedDFP avoids direct sharing of raw data, model weights, or gradients, which are more susceptible to inversion or leakage attacks [48, 49]. Although a formal privacy analysis is not included in this work, the minimal and compact nature of prompt vectors suggests a reduced risk surface compared to existing gradient-based or data-sharing approaches. As feature-level prompts do not encode detailed representations of raw images or patient identity, they offer an inherently safer abstraction for federated updates. Future work will explore incorporating differential privacy mechanisms and formal guarantees to further strengthen this aspect, laying the foundation for robust and compliant deployment in large-scale digital pathology networks

Despite its advantages, FedDFP has several limitations. The current formulation assumes that all participating centres use a compatible feature extractor, such as a shared ResNet50 or ViT backbone. While this enables consistent embedding spaces across clients, it may limit applicability in environments where feature extractor architectures vary significantly or are not centrally coordinated. Additionally, the current study does not address challenges such as client dropout, non-IID feature distributions, or adversarial client behaviour, all of which are important considerations for real-world deployment.

In future works, several important directions can further enhance the capabilities and applicability of FedDFP. First, while the current formulation assumes a shared feature extractor across institutions, future work could explore adaptations of FedDFP to settings with non-aligned or heterogeneous feature spaces, such as through domain-invariant prompt representations or cross-modal embeddings [50]. Second, integrating FedDFP with self-supervised or foundation model-based feature extractors [51, 52] could expand its utility to scenarios where labelled data is scarce or task definitions

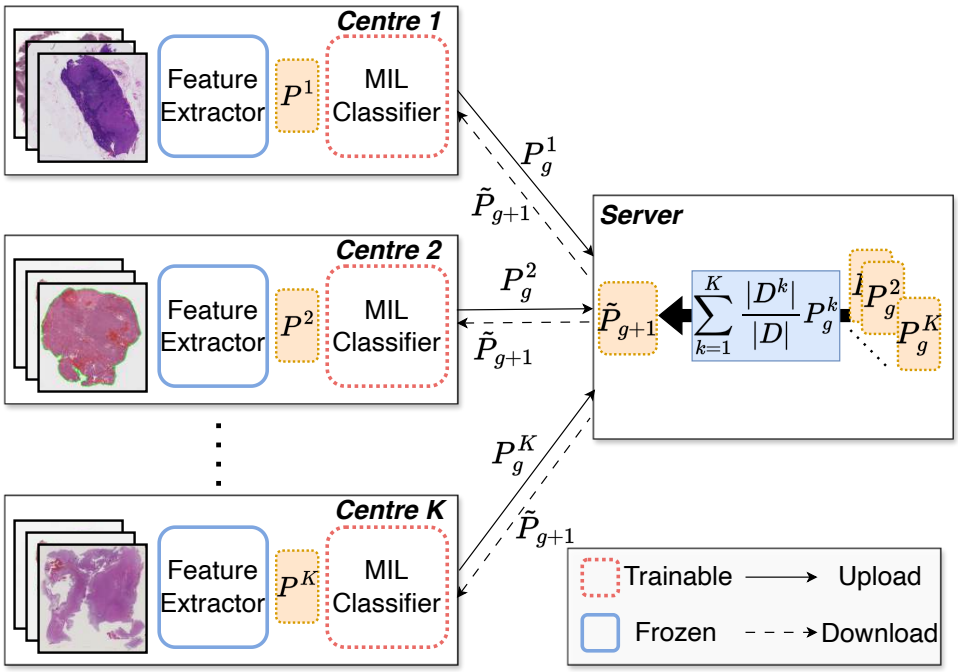


Fig. 6 Overview of FedDFP. For each medical centre, a pre-trained feature extractor is used to map raw images to feature embeddings, then these feature embeddings interact with our proposed Deep Feature Prompt (DFP, P^k), the MIL classifier then aggregates the modified feature embeddings to obtain the slide-level prediction. The server aggregates the locally learned DFPs and updates the global DFP which is downloaded by each centre for the next round of training.

vary over time. Finally, the current design does not explicitly address robustness to client dropout [53] or adversarial clients [54], all of which are prevalent in real-world federated environments; incorporating mechanisms for resilience and trustworthiness remains a crucial next step.

4 Method

4.1 Overview of the FedDFP framework

The Federated Deep Feature Prompt (FedDFP) framework was developed specifically to address key challenges in federated digital pathology, including institutional heterogeneity, communication efficiency, and privacy concerns. An overview of the FedDFP framework is illustrated in Fig. 6. FedDFP introduces a lightweight, learnable prompt vector applied directly to feature embeddings generated by pre-trained neural networks (*e.g.*, CNNs or ViT). These compact prompts, which can be exchanged between individual clients and the central aggregation server, substantially streamlined federated communication and enabled personalised modelling at each centre.

4.2 Personalised Federated WSI Classification

In this study, we address the practical challenge of slide-level WSI classification, where only slide-level labels are available, and the task is typically approached in a weakly-supervised manner [6, 9, 12]. Formally, given a WSI image x and its label y, we first crop x into a set of patches $x = \{c_i\}_{i=1}^{N}$, where N could vary across WSIs. Then a pre-trained feature extractor ($\mathcal{F}$) is used to convert c_i to patch embedding $e_i = \mathcal{F}(c_i)$. This is usually done in an offline manner where e_i is then aggregated to form a bag representation that is further handled by the MIL classifier ($\mathcal{M}$) to obtain the slide-level prediction:

$$\hat{y} = \mathcal{M}\left(\{e_i\}_{i=1}^{N}\right) \tag{1}$$

The standard Federated WSI classification involves multiple medical centres working collaboratively to enhance the performance of a global MIL model. In practical scenarios, different medical centres may have different computational resources and wish to adopt different feature extractors or MIL classifiers. Thus, we focus on the personalised federated WSI classification to tackle this heterogeneous local model scenario. Supposing we have K centres where each consists of their own data (*i.e.*, $\{D^1, D^2, \cdots, D^K\}$) and the task can be formulated as:

$$\underset{\{\mathcal{M}^k\}_{k=1}^{K}}{\arg\min}\ \mathcal{L} = \sum_{k=1}^{K} \frac{|D^k|}{|D|} \mathbb{E}_{(x,y)\in D^k} \mathcal{L}(\mathcal{M}^k; x, y) \tag{2}$$

Here D represents the combined dataset from all centres, $\mathcal{L}$ is a task-specific loss function, and different MIL methods define $\mathcal{L}$ in various forms. For example, in CLAM [10], $\mathcal{L}$ comprises both slide-level and patch-level loss components. Moreover, $\mathcal{M}^k$ denotes the MIL classifier selected by centre k. If all centres choose to use the same MIL classifier, this scenario becomes a homogeneous local model setup, and standard FL methods can be applied.

4.3 Federated Deep Feature Prompt Learning

The overall framework of FedDFP is shown in Fig. 6. At each local centre k, a pre-trained feature extractor is used for mapping the raw patch images to patch feature embeddings. However, domain discrepancies between the generic images used to train these feature extractors and the specific WSI dataset in use may lead to suboptimal feature representations. Thus, the proposed DFP is used to directly interact with feature representations for task-specific feature extraction without changing the pre-trained feature extractor. Then the updated feature representations are used to train the downstream MIL classifier (M^k).

Specifically, supposing we have G communication rounds, for each g round, each participating centre k conducts its local training and obtains the optimal P_g^k and M_g^k (see Section 4.4 for details). During communication, the trained M_g^k is kept locally, while P_g^k is uploaded to the server. Due to the parameter-efficient nature of P_g^k, uploading P_g^k leads to much lower communication costs compared to uploading M_g^k. This effectively alleviates the issue of limited communication bandwidth. On the server end,

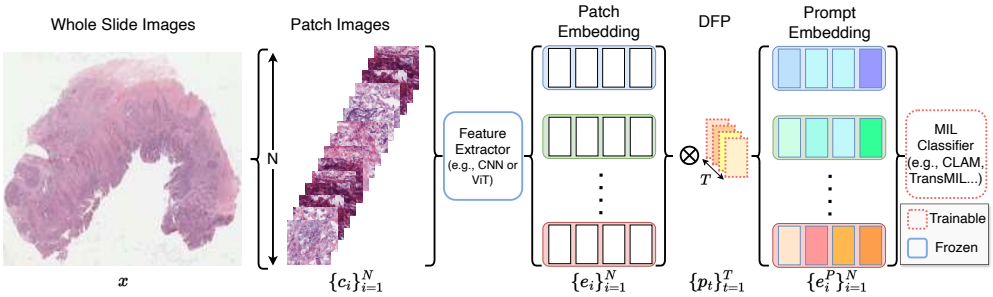


Fig. 7 The workflow of Deep Feature Prompt in a local centre. Instead of applying visual prompts to the images, we apply visual prompts to their features. The prompted features are then used as inputs to train the MIL classifier for downstream tasks.

the trained DFPs from each centre are collected and aggregated to obtain the global DFP $\tilde{P}_{g+1}$ using:

$$\tilde{P}_{g+1} = \sum_{k=1}^{K} \frac{|D^k|}{|D|} P_g^k \tag{3}$$

$\tilde{P}_{g+1}$ is then sent back to each participating centre for the next training round. 1 shows the detailed procedure of FedDFP.

4.4 Local Update with Deep Feature Prompt

Previous image-level visual prompts [55] face two challenges when applying to WSI classification: 1) a lack of uniform design and 2) being computationally extensive. Our DFP approach addresses these issues by operating directly on feature embeddings. Although CNN-based and ViT-based feature extractors differ in input formats (raw images for CNNs and tokenised images for ViTs), they ultimately produce similar one-dimensional feature vectors of varying dimensions. By implementing prompts at the feature embedding level, we establish a unified design across models, requiring only dimensional adjustments to accommodate different extractors. Furthermore, as DFP preserves the original dimension of feature embeddings (*i.e.*, inputs to MIL), it allows for flexibility in using various MIL models without modifying the prompt design. Additionally, DFP is computationally efficient as it eliminates the need to load the feature extractor during optimisation.

As illustrated in Fig. 7, given a slide (x^k) in centre k, which contains $\{c_i^k\}_{i=1}^N$ patches, we use a pre-trained $\mathcal{F}^k$ to convert each image patch $c_i^k \in \mathcal{R}^{H\times W\times 3}$ into a one-dimensional feature embedding $e_i^k = \mathcal{F}^k(c_i^k)$, where $e_i^k \in \mathcal{R}^{1\times L}$ and L is the embedding length. We then initialise T deep feature prompts $P^k = \{p_t^k\}_{t=1}^T$, where each p_t^k is of the same size as e_i^k. e_i^k is then prompted to obtain task-specific feature embedding $e_i^{k,P}$ using the following equation:

$$e_i^{k,P} = \frac{1}{T}\sum_{t=1}^{T} e_i^k \odot p_t^k \tag{4}$$

Algorithm 1: FedDFP.

Data: Training set from K medical centres $\{D^1, D^2, \cdots, D^K\}$; Global communication rounds G; Local training round Z; DFPs $\{P^1, P^2, \cdots, P^K\}$; MIL classifiers $\{\mathcal{M}^1, \mathcal{M}^2, \cdots, \mathcal{M}^K\}$; MIL learning rate $\eta_{\mathcal{M}}$; DFP learning rate η_P

/* ServerUpdate ***/**

Randomly initialise a global DFP $\tilde{P}_g$;

for $g = 0, \cdots, G-1$ **do**

  for $k = 0, \cdots, K-1$ **do**

    $P^k \leftarrow \tilde{P}_g$;

    $P_g^k \leftarrow$ **LocalUpdate**$(P^k, \mathcal{M}^k, D^k)$;

  end

  $\tilde{P}_{g+1} \leftarrow \sum_{k=1}^{K} \frac{|D^k|}{|D|} P_g^k$;

end

return $\tilde{P}_{g+1}$;

/* LocalUpdate ***/**

$P^{k,z=0} \leftarrow P^k$;

$M^{k,z=0} \leftarrow M^k$;

for $z = 0, \cdots, Z-1$ **do**

  Sample (x, y) from D^k;

  Update using Equation 6; $P^{k,z+1} \leftarrow P_g^{k,z} - \eta_P \nabla \mathcal{L}(\mathcal{M}^k, P^{k,z}; x, y)$;

  $\mathcal{M}^{k,z+1} \leftarrow \mathcal{M}^{k,z} - \eta_{\mathcal{M}} \nabla \mathcal{L}(\mathcal{M}^{k,z}, P^{k,z}; x, y)$;

end

return $P^{k,Z}$;

where element-wise multiplication is applied between the feature embedding and the deep feature prompt. Once we have the prompted feature embedding of each image patch, we can concatenate them to form $\{e_i^{k,P}\}_{i=1}^N$. This aggregated feature set is then used as input to an MIL classifier to obtain the slide-level prediction:

$$\hat{y}^k = \mathcal{M}^k \left(\{e_i^{k,P}\}_{i=1}^N \right) \tag{5}$$

In practice, each prompt is optimised together with the downstream MIL classifier M^k without loading $\mathcal{F}^k$ using the following equation:

$$\underset{\{\mathcal{M}^k\}_{k=1}^K, \{P^k\}_{k=1}^K}{\arg\min} \mathcal{L} = \sum_{k=1}^{K} \frac{|D^k|}{|D|} \mathbb{E}_{(x,y)\in D^k} \mathcal{L}(\mathcal{M}^k, P^k; x, y) \tag{6}$$

Besides being lightweight and compatible with various feature extractors and MIL classifiers, DFP, a feature-space visual prompt, typically induces lower variability in the feature space. This, in turn, leads to tighter generalisation bounds on the expected

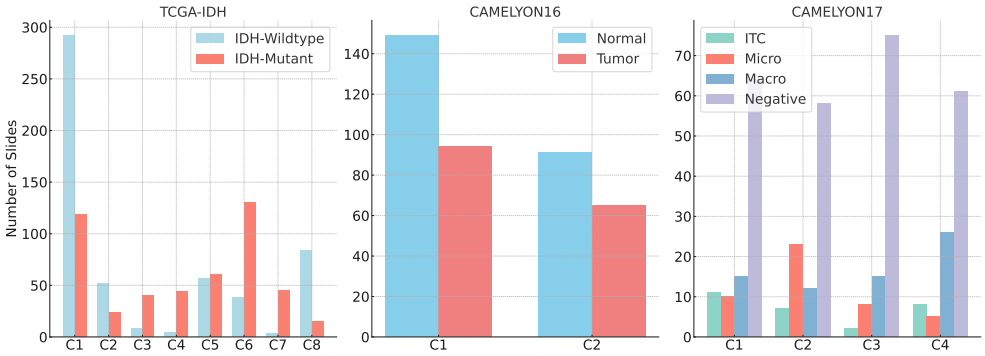


Fig. 8 Distribution of slides across participating centres in the three datasets.

loss, suggesting that feature-level prompts may help the downstream MIL classifier generalise more effectively.

4.5 Dataset Overview & Slide Distribution

The experiments were conducted using three public datasets, TCGA-IDH, CAMELYON16 and CAMELYON17, covering two major disease types, including brain and breast cancer. For TCGA-IDH, we included sites contributing over 40 slides, yielding a total of 968 slides across 8 centres (C1-C8). Each slide is labelled as either IDH wildtype (WT) or mutant (MU) based on immunohistochemistry and/or genetic sequencing. Centre-specific details are as follows: C1: 292 WT / 119 MU, C2: 52 WT / 24 MU, C3: 8 WT / 40 MU, C4: 4 WT / 44 MU, C5: 57 WT / 61 MU, C6: 38 WT / 130 MU, C7: 3 WT / 45 MU, and C8: 84 WT / 15 MU. Within each centre, we randomly allocated 80% of WSIs to training and 20% to testing. For CAMELYON16, the task is to classify slides as either normal or tumour. The dataset includes 399 WSIs from two medical centres: RUMC (C1), with 169 slides for training (99 normal / 70 tumour) and 74 slides for testing (50 normal / 24 tumour); and UMCU (C2), contributing 101 slides for training (60 normal / 41 tumour) and 55 slides for testing (31 normal / 24 tumour). For CAMELYON17, the task is to classify each slide into one of four categories: negative (neg), isolated tumour cells (itc), micro-metastases (micro), and macro-metastases (macro). Each centre (C1–C5) includes 20 patients, with five slides per patient. The class distribution across centres is as follows: C1: 11 itc, 10 micro, 15 macro, and 64 neg; C2: 7 itc, 23 micro, 12 macro, and 58 neg; C3: 2 itc, 8 micro, 15 macro, and 75 neg; C4: 8 itc, 5 micro, 26 macro, and 61 neg. Within each centre, we randomly allocated 80% of the data (slides from 16 patients) for training and 20% for testing (slides from 4 patients), ensuring that slides from the same patient were assigned to the same split. The detailed slide distributions across different centres are shown in Fig. 8.

4.6 Data Preprocessing

For all datasets, we adopted the preprocessing pipeline established by CLAM [10] to ensure consistency with prior work in weakly-supervised whole slide image (WSI) classification. WSIs were first segmented to exclude background and artifacts using colour deconvolution and morphological operations. Patch-level sampling was then performed based on dataset-specific magnifications and resolutions. For TCGA-IDH [56], slides were loaded at 10× magnification and divided into non-overlapping 1024×1024 pixel patches. For CAMELYON17 [57] and CAMELYON16 [58], slides were processed at 20× and 40× magnification, respectively, with 256×256 patches extracted from tissue regions. All extracted patches were resized to 224×224 pixels prior to feature extraction.

Patch-level features were obtained using pre-trained models, including ResNet50 and ViT-S/16, which are both trained on ImageNet [59]. These models generated fixed-length embeddings, which served as the input to downstream multiple instance learning (MIL) classifiers. Each dataset was labelled at the slide level for supervised classification: TCGA-IDH slides were labelled as IDH mutant or wildtype; CAMELYON16 slides were labelled as tumour or normal; and CAMELYON17 included four classes—negative, isolated tumour cells (ITC), micro-metastases, and macro-metastases. Train-test splits were generated independently for each institution: For TCGA-IDH, institutions contributing at least 40 slides were retained. Within each institution, 80% of slides were randomly assigned to the training set and 20% to the test set. For CAMELYON16, we followed the predefined training and test splits from the original challenge, maintaining institution-specific separation. For CAMELYON17, as official test labels are not publicly available, we used the entire training set (500 slides from 100 patients). Within each institution, we performed an 80/20 split at the patient level, ensuring that all slides from a single patient were assigned to either the training or test set to prevent data leakage. All preprocessing, patch feature extraction were conducted locally per institution, with no exchange of raw WSIs or patch-level data, faithfully reflecting federated learning constraints.

4.7 Training Details

All models were trained under the federated learning (FL) setting, with no data exchanged across clients. For each dataset, both WSIs preprocessing and feature embedding extraction were performed independently at each centre. Feature extractors remained frozen throughout training to simulate common clinical constraints and reduce computational demands.

For the FL setup, we employed 100 global communication rounds ($G = 100$), with each client conducting 50 local training rounds ($Z = 50$) between global updates. Both local DFPs and downstream MIL classifiers were trained using the Adam optimiser with a learning rate of 0.0003. DFPs were randomly initialised and, unless otherwise specified, a single prompt ($T = 1$) was learned per client to minimise communication overhead.

All experiments were conducted on a single NVIDIA V100 GPU. The evaluation was performed using both per-centre and globally averaged test accuracy. Additionally, we reported the micro-averaged AUC and ROC curve, averaged across five independent runs with fixed data splits to assess the impact of random initialisation. Standard deviations are reported alongside mean performance metrics to reflect training stability.

To ensure reproducibility and fair comparison, consistent random seeds and training protocols were applied across all baseline methods and FedDFP.

4.8 Baseline Methods

To evaluate the effectiveness of FedDFP, we conducted comprehensive comparisons against two categories of baselines: (1) federated learning (FL) algorithms, to assess collaborative learning performance, and (2) multiple instance learning (MIL) architectures, to evaluate FedDFP's robustness in heterogeneous model settings commonly encountered in digital pathology.

Federated Learning Baselines

We compared FedDFP to a broad range of state-of-the-art FL methods. These include both standard FL approaches utilising a globally shared model, and personalised FL methods allowing for personalised local models chosen by the clients. Standard methods include: FedHisto [13], a histopathology-specific adaptation of FedAvg; FedDyn [41], which uses dynamic regularization to mitigate client drift; MOON [42], which employs contrastive loss to align local and global models; and SCAFFOLD [43], which uses control variates to correct for client heterogeneity. Personalised FL baselines include: FedProto [35], which transmits class prototypes to personalise local classifiers; DESA [28], which facilitates knowledge sharing via synthetic data generation; SGPT [31], a spatial visual prompt learning approach; and pFedPG [37], a prompt-based gradient optimisation framework for personalisation.

Multiple Instance Learning Frameworks

Given the nature of WSI analysis in digital pathology, we also evaluated FedDFP in a heterogeneous federated setting where clients may adopt different MIL architectures. This setup reflects realistic deployment environments in which clinical centres select models based on their computational constraints and workflow preferences. We compared against three widely used MIL classifiers: ABMIL [45], which uses attention-based pooling to aggregate patch-level features; CLAM [10], which introduces clustering-based attention for better instance discrimination; and TransMIL [44], a transformer-based model that captures inter-patch dependencies. These models were randomly assigned to different institutions in our heterogeneous experimental setup, allowing us to evaluate FedDFP's compatibility and performance under variable local model architectures.

5 Data Availability

All datasets used in this study are publicly available and have been previously published. The TCGA-IDH dataset was obtained from The Cancer Genome Atlas via the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). The CAMELYON16 and CAMELYON17 datasets are available through the CAMELYON Grand Challenge website (https://camelyon16.grand-challenge.org/ and https://camelyon17.grand-challenge.org/,respectively). No additional restrictions apply to access or use of these datasets. All code and scripts used for data preprocessing and training are available upon reasonable request and will be made public upon publication.

6 Code Availability

The code used to implement FedDFP, along with scripts for data preprocessing, model training, and evaluation, will be made publicly available on GitHub (https://github.com/thomascong121/FedDFP) upon publication of this manuscript. This includes configurations for all baseline methods to ensure reproducibility and facilitate benchmarking. Until then, the code is available from the corresponding author upon reasonable request.

7 Acknowledgements

This work was supported by the NHMRC Centre of Research Excellence in Digital Health, an NHMRC Ideas Grant (GNT202035) and an NHMRC Investigator Grant (GNT2008645).

References

[1] Jahn, S.W., Plass, M., Moinfar, F.: Digital pathology: advantages, limitations and emerging perspectives. Journal of clinical medicine **9**(11), 3697 (2020)

[2] Baxi, V., Edwards, R., Montalto, M., Saha, S.: Digital pathology and artificial intelligence in translational medicine and clinical practice. Modern Pathology **35**(1), 23–32 (2022)

[3] Cong, C., Liu, S., Di Ieva, A., Pagnucco, M., Berkovsky, S., Song, Y.: Colour adaptive generative networks for stain normalisation of histopathology images. Medical Image Analysis **82**, 102580 (2022)

[4] Waqas, A., Bui, M.M., Glassy, E.F., El Naqa, I., Borkowski, P., Borkowski, A.A., Rasool, G.: Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models. Laboratory investigation **103**(11), 100255 (2023)

[5] Jose, L., Liu, S., Russo, C., Cong, C., Song, Y., Rodriguez, M., Di Ieva, A.: Artificial intelligence–assisted classification of gliomas using whole slide images. Archives of Pathology & Laboratory Medicine **147**(8), 916–924 (2023)

[6] Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)

[7] Di Ieva, A., Bruner, E., Widhalm, G., Minchev, G., Tschabitscher, M., Grizzi, F.: Computer-assisted and fractal-based morphometric assessment of microvascularity in histological specimens of gliomas. Scientific Reports **2**(1), 429 (2012)

[8] Chikontwe, P., Kim, M., Nam, S.J., Go, H., Park, S.H.: Multiple instance learning with center embeddings for histopathology classification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23, pp. 519–528 (2020). Springer

[9] Qu, L., Luo, X., Liu, S., Wang, M., Song, Z.: Dgmil: Distribution guided multiple instance learning for whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 24–34 (2022). Springer

[10] Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)

[11] Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14318–14328 (2021)

[12] Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16144–16155 (2022)

[13] Lu, M.Y., Chen, R.J., Kong, D., Lipkova, J., Singh, R., Williamson, D.F., Chen, T.Y., Mahmood, F.: Federated learning for computational pathology on gigapixel whole slide images. Medical image analysis **76**, 102298 (2022)

[14] Wirth, F.N., Meurers, T., Johns, M., Prasser, F.: Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. BMC Medical Informatics and Decision Making **21**, 1–13 (2021)

[15] Feng, C.-M., Li, B., Xu, X., Liu, Y., Fu, H., Zuo, W.: Learning federated visual prompt in null space for mri reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8064–8073 (2023)

[16] Pati, S., Kumar, S., Varma, A., Edwards, B., Lu, C., Qu, L., Wang, J.J., Lakshminarayanan, A., Wang, S.-h., Sheller, M.J., et al.: Privacy preservation for federated learning in health care. Patterns **5**(7) (2024)

[17] Huang, L., Shao, L., Bao, M., Guo, C., Shao, Z., Huang, X., Wang, M., Jiang, X., Hu, S.: Federated learning with comparative learning-based dynamic parameter updating on glioma whole slide images. Engineering Applications of Artificial Intelligence **137**, 109233 (2024)

[18] Terrail, J., Leopold, A., Joly, C., Béguier, C., Andreux, M., Maussion, C., Schmauch, B., Tramel, E.W., Bendjebbar, E., Zaslavskiy, M., *et al.*: Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. Nature medicine **29**(1), 135–146 (2023)

[19] Andreux, M., Terrail, J.O., Beguier, C., Tramel, E.W.: Siloed federated learning for multi-centric histopathology datasets. In: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2, pp. 129–139 (2020). Springer

[20] Zhang, Y., Li, Z., Han, X., Ding, S., Li, J., Wang, J., Ying, S., Shi, J.: Pseudo-data based self-supervised federated learning for classification of histopathological images. IEEE Transactions on Medical Imaging (2023)

[21] McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282 (2017). PMLR

[22] Zhang, F., Zhai, D., Bai, G., Jiang, J., Ye, Q., Ji, X., Liu, X.: Towards fairness-aware and privacy-preserving enhanced collaborative learning for healthcare. Nature Communications **16** (2025) https://doi.org/10.1038/s41467-025-58055-3

[23] Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., Qadir, J.: Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine **158**, 106848 (2023)

[24] Boenisch, F., Dziedzic, A., Schuster, R., Shamsabadi, A.S., Shumailov, I., Papernot, N.: When the curious abandon honesty: Federated learning is not private. In: 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P), pp. 175–199 (2023). IEEE

[25] Arivazhagan, M.G., Aggarwal, V., Singh, A.K., Choudhary, S.: Federated learning with personalization layers. arXiv preprint arXiv:1912.00818 (2019)

[26] Collins, L., Hassani, H., Mokhtari, A., Shakkottai, S.: Exploiting shared representations for personalized federated learning. In: International Conference on

Machine Learning, pp. 2089–2099 (2021). PMLR

[27] Wu, C., Wu, F., Lyu, L., Huang, Y., Xie, X.: Communication-efficient federated learning via knowledge distillation. Nature communications **13**(1), 2032 (2022)

[28] Huang, C.-Y., Srinivas, K., Zhang, X., Li, X.: Overcoming data and model heterogeneities in decentralized federated learning via synthetic anchors. arXiv preprint arXiv:2405.11525 (2024)

[29] Xiong, Y., Wang, R., Cheng, M., Yu, F., Hsieh, C.-J.: Feddm: Iterative distribution matching for communication-efficient federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16323–16332 (2023)

[30] Wang, Y., Fu, H., Kanagavelu, R., Wei, Q., Liu, Y., Goh, R.S.M.: An aggregation-free federated learning for tackling data heterogeneity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26233–26242 (2024)

[31] Jia, Y., Vahidian, S., Sun, J., Zhang, J., Kungurtsev, V., Gong, N.Z., Chen, Y.: Unlocking the potential of federated learning: The symphony of dataset distillation via deep generative latents. In: European Conference on Computer Vision, pp. 18–33 (2025). Springer

[32] Jin, H., Liu, S., Cong, C., Feng, Q., Liu, Y., Huang, L., Hu, Y.: Fedwsidd: Federated whole slide image classification via dataset distillation. arXiv preprint arXiv:2506.15365 (2025)

[33] Jeong, E., Oh, S., Kim, H., Park, J., Bennis, M., Kim, S.-L.: Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. arXiv preprint arXiv:1811.11479 (2018)

[34] Chan, Y.H., Ngai, E.C.: Fedhe: Heterogeneous models and communication-efficient federated learning. In: 2021 17th International Conference on Mobility, Sensing and Networking (MSN), pp. 207–214 (2021). IEEE

[35] Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., Zhang, C.: Fedproto: Federated prototype learning across heterogeneous clients. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 8432–8440 (2022)

[36] Li, G., Wu, W., Sun, Y., Shen, L., Wu, B., Tao, D.: Visual prompt based personalized federated learning. arXiv preprint arXiv:2303.08678 (2023)

[37] Yang, F.-E., Wang, C.-Y., Wang, Y.-C.F.: Efficient model personalization in federated learning via client-specific prompt generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 19159–19168 (2023)

[38] Deng, W., Thrampoulidis, C., Li, X.: Unlocking the potential of prompt-tuning in bridging generalized and personalized federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6087–6097 (2024)

[39] He, K., Zhang, X., *et al.*: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)

[40] 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. arXiv preprint arXiv:2010.11929 (2020)

[41] Acar, D.A.E., Zhao, Y., Navarro, R.M., Mattina, M., Whatmough, P.N., Saligrama, V.: Federated learning based on dynamic regularization. arXiv preprint arXiv:2111.04263 (2021)

[42] Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10713–10722 (2021)

[43] Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. In: International Conference on Machine Learning, pp. 5132–5143 (2020). PMLR

[44] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., *et al.*: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)

[45] Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136 (2018). PMLR

[46] Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., Socher, R.: Deep learning-enabled medical computer vision. NPJ digital medicine **4**(1), 5 (2021)

[47] Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature medicine **25**(1), 44–56 (2019)

[48] Hatamizadeh, A., Yin, H., Molchanov, P., Myronenko, A., Li, W., Dogra, P., Feng, A., Flores, M.G., Kautz, J., Xu, D., *et al.*: Do gradient inversion attacks make federated learning unsafe? IEEE Transactions on Medical Imaging **42**(7), 2044–2056 (2023)

[49] Feng, S., Tramèr, F.: Privacy backdoors: stealing data with corrupted pretrained

models. arXiv preprint arXiv:2404.00473 (2024)

[50] Zhao, C., Wang, Y., Jiang, X., Shen, Y., Song, K., Li, D., Miao, D.: Learning domain invariant prompt for vision-language models. IEEE Transactions on Image Processing **33**, 1348–1360 (2024)

[51] Zhang, J., Kapse, S., Ma, K., Prasanna, P., Saltz, J., Vakalopoulou, M., Samaras, D.: Prompt-mil: Boosting multi-instance learning schemes via task-specific prompt tuning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 624–634 (2023). Springer

[52] Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., *et al.*: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024)

[53] Wang, H., Xu, J.: Combating client dropout in federated learning via friend model substitution. arXiv preprint arXiv:2205.13222 (2022)

[54] Allouah, Y., El Mrini, A., Guerraoui, R., Gupta, N., Pinot, R.: Fine-tuning personalization in federated learning to mitigate adversarial clients. Advances in Neural Information Processing Systems **37**, 100816–100844 (2024)

[55] Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274 (2022)

[56] Liu, S., Shah, Z., Sav, A., Russo, C., Berkovsky, S., Qian, Y., Coiera, E., Di Ieva, A.: Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. Scientific Reports **10**(1), 1–11 (2020)

[57] Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., *et al.*: From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Transactions on Medical Imaging **38**(2), 550–560 (2019)

[58] Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., *et al.*: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**(22), 2199–2210 (2017)

[59] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee