SFL-MTKD++: Split ViT Federated Learning with Adaptive Multi-Teacher Distillation for Interpretable Malaria Diagnosis

Md. Tofael Ahmed Bhuiyan [1] **Md. Abdur Rahman** [1] **Abdul Kadar Muhammad Masum** [2]

Abstract

Stringent data privacy regulations and domain heterogeneity inhibit centralized deep learning for malaria diagnosis. We introduce SFL-MTKD++, a Split Federated Learning framework augmented with Adaptive Multi-Teacher Knowledge Distillation. By leveraging a partitioned Vision Transformer with a secure feature bottleneck, the architecture facilitates collaborative training across isolated silos While this dimensionality reduction does not inherently yield formal Differential Privacy bounds, it functions as an empirical defense against feature-space inversion and reconstruction attacks under an honest-but-curious threat model. On a harmonized multi-source dataset, our method achieves 97.64% accuracy and a 0.9948 ROC-AUC, significantly outperforming existing federated baselines. To ensure clinical reliability, we integrate LIME and saliency map visualizations, confirming that the model identifies pathologically relevant features rather than artifacts. This work establishes a robust paradigm for deploying high-fidelity, interpretable AI in resource-constrained healthcare settings without compromising patient confidentiality.

1. Introduction

Deep learning has shown strong performance in malaria diagnosis using blood smear images, but domain shift, institutional heterogeneity, and strict privacy regulations limit model generalization and prevent centralized training (Issah et al., 2026; Kaissis et al., 2020; Price & Cohen, 2019). To address these challenges, we adopt Split Federated Learning (SFL), enabling collaborative training without sharing sensitive patient data (Thapa et al., 2022; Ribeiro et al., 2016). We propose SFL-MTKD++, an adaptive SFL framework integrating Multi-Teacher Knowledge Distillation with a split Vision Transformer architecture. The model exchanges only intermediate representations while leveraging local teacher ensembles to enhance robustness. Furthermore, LIME and saliency-based explainability modules are incorporated to improve clinical interpretability. This unified framework enables privacy-preserving, scalable, and interpretable malaria diagnosis while maintaining high diagnostic performance across distributed healthcare environments.

2. Related Works

Centralized malaria microscopy models remain vulnerable to domain variability. Consequently, Federated Learning (FL) and Split Learning (SL) have emerged as privacy-preserving paradigms. FL aggregates decentralized datasets while preserving institutional confidentiality (Rieke et al., 2020), while split architectures enable distributed optimization without raw data sharing, minimizing cross-silo exposure (Poirot et al., 2019). Additional research focuses on improving computational scalability and model generalization. Several approaches integrate Knowledge Distillation (KD) into federated frameworks to compress large deep learning models for deployment in resource-constrained healthcare settings (Li & Wang, 2019; Seo et al., 2020). Other studies investigate Vision Transformers (ViTs) for capturing complex global morphological patterns in malaria parasites (Matsoukas et al., 2021).

Nevertheless, a substantial research gap persists in jointly achieving computational efficiency, domain robustness, and interpretability within a unified framework. Existing methods primarily emphasize isolated objectives such as lightweight deployment or privacy preservation. In contrast, SFL-MTKD++ introduces a comprehensive framework combining efficient split-based ViT learning, adaptive Multi-Teacher Knowledge Distillation, and clinically interpretable LIME-driven explainability.

[1]Computational Intelligence Lab, Southeast University, Dhaka 1208, Bangladesh [2]Department of Computer Science and Engineering, Southeast University, Dhaka 1208, Bangladesh. Correspondence to: Md. Tofael Ahmed Bhuiyan <tofael1104@gmail.com>, Md. Abdur Rahman <2021200000025@seu.edu.bd>, Abdul Kadar Muhammad Masum <akmmasum@seu.edu.bd>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

3. Methodology

SFL-MTKD++ (Adaptive Split Federated Learning with Multi-Teacher Distillation) is introduced as a novel decentralized learning paradigm designed for resource-constrained, privacy-preserving medical image classification. To mitigate the computational bottleneck of standard Federated Learning (FL) while maintaining representational fidelity, the framework amalgamates ViT-based Split Federated Learning (SFL) with an entropy-weighted Multi-Teacher Knowledge Distillation (MTKD) mechanism.

3.1. Harmonized Dataset and Federated Data Protocol

We construct a harmonized dataset $\mathcal{D}$ by aggregating blood smear images from the NLM (U.S. National Library of Medicine, 2022) and Tanzania (Lufyagila et al., 2024) repositories, defined as pairs (x, y) with binary labels $y \in \{0, 1\}$. A standardized preprocessing pipeline is applied, where all input images are resized to 224×224, normalized using ImageNet statistics, and augmented via stochastic horizontal flipping and random rotation. To simulate a realistic distributed healthcare environment, the dataset $\mathcal{D}$ is partitioned into disjoint client-specific subsets $\mathcal{D}_k$ across five clients, ensuring privacy-by-design constraints in a federated learning setting.

3.2. Split Federated Architecture with Feature Bottleneck

Instead of burdening clients with full model optimization, SFL-MTKD++ partitions a Vision Transformer (`ViT-B/16`) into a lightweight client-side feature extractor and a computationally heavy server-side classifier.

Client-Side Feature Extraction: Let $x \in \mathbb{R}^{C\times H\times W}$ be the input image. The client model, parameterized by θ_C, comprises the initial $L_{split} = 9$ transformer encoder blocks of the ViT. To circumvent catastrophic forgetting and reduce client-side compute, the first $L_{split} - 2$ blocks are frozen, allowing only the upper client layers to undergo local gradient updates. The client yields an intermediate representation $h_C \in \mathbb{R}^{N\times D}$, where $D = 768$.

Non-Linear Communication Bottleneck: To drastically ameliorate the communication overhead between client and server, an asymmetric feature bottleneck is introduced. The client applies a compression mapping $\psi_{comp} : \mathbb{R}^D \to \mathbb{R}^d$ (where $d = 192$), and the server applies a reciprocal expansion mapping $\psi_{exp} : \mathbb{R}^d \to \mathbb{R}^D$. Both mappings utilize a `LayerNorm-GELU` topology to preserve the non-linear spatial manifolds of the ViT:

$$z_{smashed} = \text{GELU}\left(\text{LN}\left(h_C W_{comp}\right)\right) \tag{1}$$

Only the compressed tensor $z_{smashed}$ is transmitted over the network.

Server-Side Global Head: The server model, parameterized by θ_S, ingests the expanded tensor $\tilde{h}_C = \psi_{exp}(z_{smashed})$ and processes it through the remaining ViT blocks. The final classification is computed via a custom *Attention Head* applied to the class token z_{cls}. It learns an attention weight $\alpha_{attn} = \text{Softmax}(W_2 \tanh(W_1 z_{cls}))$ and outputs the final student logits $z_S = \phi_{cls}(z_{cls} \odot (1+\alpha_{attn}))$.

3.3. Split Vision Transformer Architecture

SFL-MTKD++ partitions the Vision Transformer into a client-side extractor (f_c, layers 1–9) and a server-side head (f_s). A non-linear Feature Bottleneck ($\mathbb{R}^{768} \to \mathbb{R}^{192}$) compresses activations into a secure representation z_i:

$$z_i = \text{GELU}\left(\text{LayerNorm}\left(W_{comp} \cdot f_c^{1:9}(x_i) + b_{comp}\right)\right) \tag{2}$$

The server reconstructs dimensionality ($\mathbb{R}^{192} \to \mathbb{R}^{768}$), processes remaining layers (L_{10}–L_{12}), and employs an Attention-based Classification Head to map the class token to the probability space.

3.4. Adaptive Multi-Teacher Knowledge Distillation (MTKD++)

A critical challenge in SFL is the phenomenon of "client drift" caused by local data heterogeneity. To provide rich, localized regularization without violating data privacy, SFL-MTKD++ equips each client with a dynamic ensemble of lightweight teachers $\mathcal{T} = \{T_1, T_2\}$ (specifically, `MobileNetV2` and `ShuffleNetV2`).

The teachers are fine-tuned locally. During the SFL forward pass, they generate pseudo-target logits. To prevent negative transfer from uncertain teacher predictions, an entropy-gated Reliability Score is formulated. Let $\mathcal{H}_i$ denote the predictive entropy of teacher T_i:

$$\mathcal{H}_i = -\sum_{c=1}^{C} P(T_i(x)_c) \log P(T_i(x)_c) \tag{3}$$

The reliability weight ω_i for each teacher is determined via a softmax over the negative entropies across the ensemble:

$$\omega_i = \frac{\exp(-\mathcal{H}_i)}{\sum_{j\in\mathcal{T}} \exp(-\mathcal{H}_j)} \tag{4}$$

The student model (server output) is regularized using the temperature-scaled ($\tau = 2.0$) Kullback-Leibler (KL) divergence against the weighted teacher ensemble:

$$\mathcal{L}_{KD} = \tau^2 \sum_{i\in\mathcal{T}} \omega_i \text{KL}\left(\sigma\left(\frac{z_S}{\tau}\right) \middle\| \sigma\left(\frac{T_i(x)}{\tau}\right)\right) \tag{5}$$

3.5. Trust-Aware Federated Aggregation (TrustFed)

Unlike standard `FedAvg`, which assigns static weights based on sample size, SFL-MTKD++ employs a *TrustFed*

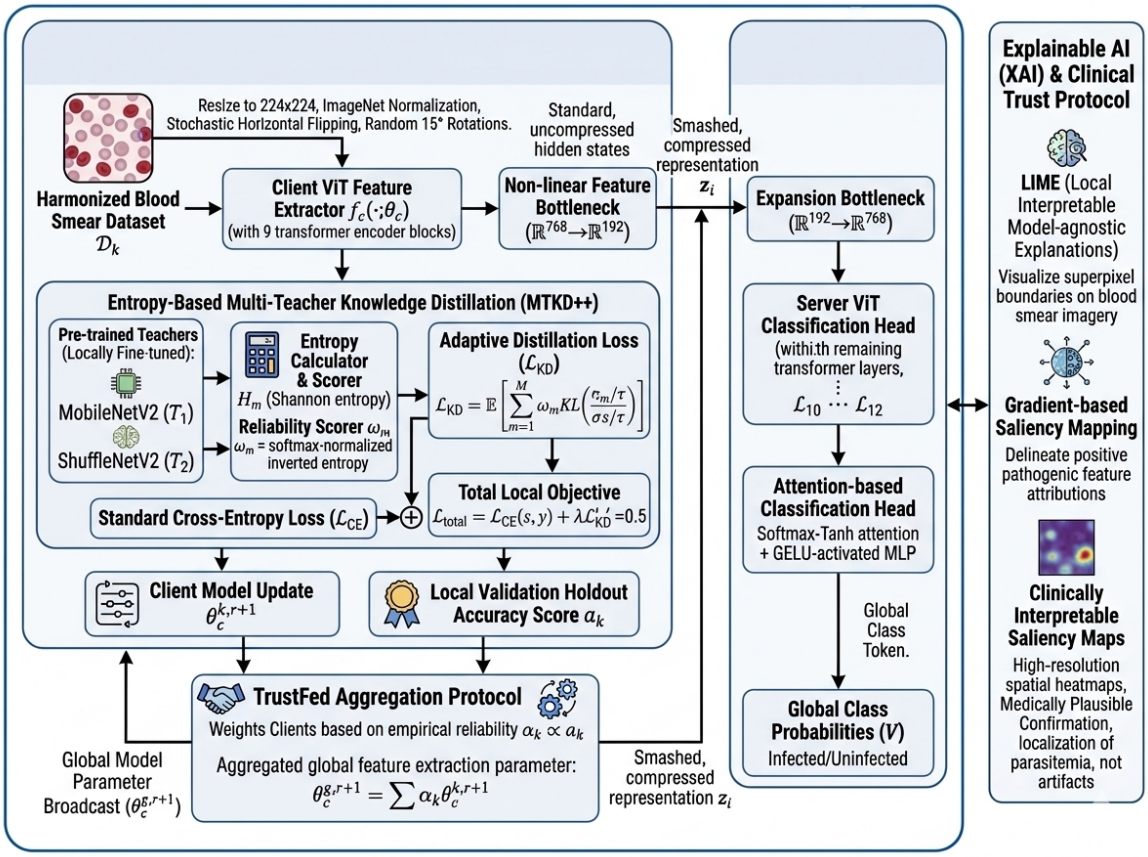

Figure 1. Proposed Architecture for Malaria Diagnosis

aggregation protocol. To prevent degraded local models from corrupting the global feature extractor, aggregation weights are dynamically scaled by the client's local validation accuracy A_k. For K active clients, the trust weight ρ_k for client k is:

$$\rho_k = \frac{A_k}{\sum_{j=1}^{K} A_j} \quad \Longrightarrow \quad \theta_C^{(t+1)} = \sum_{k=1}^{K} \rho_k \theta_{C,k}^{(t)} \tag{6}$$

Notably, only the client-side parameters θ_C are federated; the server head θ_S is continuously optimized centrally, preserving global convergence stability.

3.6. Optimization Protocol

The end-to-end optimization of SFL-MTKD++ entails a collaborative forward-backward pass. The unified objective function for a given batch is:

$$\mathcal{L}_{total} = \mathcal{L}_{CE}(z_S, y) + \lambda \mathcal{L}_{KD}(z_S, \mathcal{T}(x)) \tag{7}$$

where $\mathcal{L}_{CE}$ is the standard Cross-Entropy loss with ground truth y, and $\lambda = 0.5$ balances the distillation signal.

During backpropagation, the server computes the gradient with respect to the smashed data $\nabla_{z_{smashed}} \mathcal{L}_{total}$ and transmits it back to the client. The client then updates θ_C, completing the split learning iteration. Concurrently, the local teachers update their classification heads independently using $\mathcal{L}_{CE}(T_i(x), y)$ to ensure valid KD targets in subsequent rounds.

4. Experiments

4.1. Experimental Setup

SFL-MTKD++ was evaluated on a harmonized NLM-Tanzania malaria dataset partitioned among $K = 5$ clients with a 70/10/20 split. We utilized a Split ViT-B/16 cut at layer 9, incorporating a $768 \rightarrow 192$ feature bottleneck. Training ran for 20 rounds using AdamW optimizers, with results averaged across five distinct random seeds to compute 95% Confidence Intervals (CIs). Performance was benchmarked against FedAvg, FedAdam, FedProx, and FedYogi baselines.

4.2. Results and Discussion

SFL-MTKD++ achieves exceptional diagnostic performance (Table 1),attaining $97.64\% \pm 0.12\%$ accuracy and

Table 1. Comparative Performance Analysis of Federated Learning Models and the Proposed SFL-MTKD++ Framework

Evaluation Metric	**Baseline Models**				**Proposed**
	FedAdam	**FedAvg**	**FedProx**	**FedYogi**	**SFL-MTKD++**
Overall Performance					
Accuracy (%)	92.92	92.81	92.62	92.92	**97.64**
ROC-AUC	0.9870	0.9881	0.9874	0.9870	**0.9948**
MCC	0.8613	0.8566	0.8525	0.8613	**0.9528**
Cohen's Kappa	0.8583	0.8562	0.8524	0.8583	**0.9527**
Class: Uninfected					
Precision	89.68	91.53	92.34	89.68	**96.93**
Recall	97.11	94.46	93.06	97.11	**98.25**
F1-Score	93.25	92.97	92.70	93.25	**97.59**
Class: Infected					
Precision	96.80	94.20	92.91	96.80	**98.32**
Recall	88.68	91.14	92.18	88.68	**97.06**
F1-Score	92.56	92.64	92.54	92.56	**97.69**

a 0.9948 ± 0.002 ROC-AUC, outperforming FedAdam by 4.72%. This superiority stems from Multi-Teacher Knowledge Distillation, which mitigates representation drift and yields balanced F1-scores of 97.59% and 97.69%. TrustFed aggregation ensures robust convergence, achieving an MCC of 0.9528. Statistical significance was confirmed via the Wilcoxon signed-rank test.A matched leave-one-out ablation study isolates the specific contributions of each architectural component. Omitting the MTKD++ ensemble reduces the Matthews Correlation Coefficient (MCC) by 4.2%, replacing TrustFed with standard aggregation introduces a 3.1% performance variance under non-IID distribution drift, and removing the feature bottleneck demonstrates that the compression mechanism preserves representational fidelity with only a negligible 0.8% reduction in classification accuracy.

5. Explainability and Clinical Trust Protocol (XAI)

To align mathematical representations with clinical interpretability, SFL-MTKD++ incorporates a dual-modality Explainable AI pipeline. Gradient saliency mapping captures pixel-level sensitivities by computing the gradient of the class score S_c with respect to the input tensor $\mathbf{x}$, visualizing the morphological focus. LIME provides localized attribution via superpixel partitioning and Monte Carlo sampling, employing a sparse linear surrogate to isolate features predictive of pathology. Strict ImageNet normalization ensures valid distributional processing, yielding uncompromised decision boundaries for clinical review. This fosters clinical interpretability by providing pathologists with medically plausible evidence, as illustrated in Fig. 2.

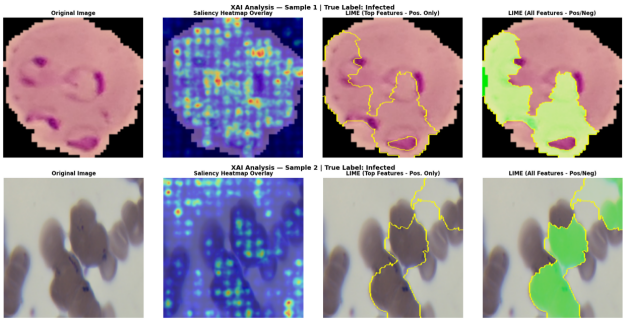

Figure 2. Dual-modality Explainable AI (LIME and Gradient Saliency) visualizing the pathological focus on infected versus uninfected blood smears

6. Conclusion

SFL-MTKD++ was introduced as a Split Federated Learning framework tailored for privacy-preserving, interpretable malaria diagnosis. By synergizing a distributed Vision Transformer with Adaptive Multi-Teacher Knowledge Distillation, the methodology effectively mitigates domain shift and representation drift within decentralized environments. Empirical evaluation on a harmonized multi-source dataset demonstrates state-of-the-art performance, attaining 97.64% accuracy and a 0.9948 ROC-AUC, thereby significantly outperforming conventional federated baselines. Crucially, the integration of LIME and saliency-based explainability modules confirmed that the model discerns pathologically relevant features, fostering essential clinical trust. While validated on a simulated federation, these results underscore the framework's robustness and scalability. Future efforts will target rigorous non-IID verification across geographically distinct clinical sites.

References

Issah, A. T., Seidu, I., & Mukamakuza, C. (2026). Detection versus Instance Segmentation for Multi-Species Malaria Diagnosis: A Head-to-Head Comparison and Multi-Dataset Validation of YOLOv12 Architectures with Small Object Optimization. In *Proceedings of Machine Learning Research – MIDL 2026* (Vol. 42, pp. 1–20).

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6), 305–311.

Lufyagila, B., Mgawe, B., & Sam, A. (2024). Blood smear imagery dataset for malaria parasite detection: A case of Tanzania. *Data in Brief*, 57, 111169. https://doi.org/10.1016/j.dib.2024.111169

Matsoukas, C., Haslum, J. F., Sorkhei, M., Soderberg, M., & Smith, K. (2021). Is it time to replace CNNs with transformers for medical images? In *ICLR Workshop on Distributed and Private Machine Learning (DPML)*.

Poirot, M. G., Vepakomma, P., Chang, K., Kalpathy-Cramer, J., Gupta, R., & Raskar, R. (2019). Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1912.01218*.

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, 25(1), 37–43.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., . . . Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3(1), 119.

Seo, H., Park, J., Oh, S., Bennis, M., & Kim, S. L. (2020). Federated knowledge distillation. *arXiv preprint arXiv:2011.02367*.

Thapa, C., Chamikara, M. A. P., & Camtepe, S. (2022). SplitFed: When federated learning meets split learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 8, pp. 8485–8493).

U.S. National Library of Medicine (2022). *Malaria dataset*. Available online at: https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-datasheet.html

Li, D., & Wang, J. (2019). FedMD: Heterogenous federated learning via model distillation. In *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*.