Privacy-Preserving Federated Vision Transformer Learning Leveraging Lightweight Homomorphic Encryption in Medical AI

AL AMIN*, Kamrul Hasan*, Liang Hong*, and Sharif Ullah†
*Department of Electrical and Computer Engineering, Tennessee State University, Nashville, TN, USA
†University of Central Arkansas, Conway, AR, USA
Email: {aamin2, mhasan1, lhong}@tnstate.edu; mullah@uca.edu

***Abstract*—Collaborative machine learning across healthcare institutions promises improved diagnostic accuracy by leveraging diverse datasets, yet privacy regulations such as HIPAA prohibit direct patient data sharing. While federated learning (FL) enables decentralized training without raw data exchange, recent studies demonstrate that model gradients in conventional FL remain vulnerable to reconstruction attacks, potentially exposing sensitive medical information. This paper presents a privacy-preserving federated learning framework combining Vision Transformers (ViT) with homomorphic encryption (HE) for secure multi-institutional histopathology classification. This approach leverages the ViT's `[CLS]` token as a compact 768-dimensional feature representation as the unit of secure aggregation, encrypting these tokens using CKKS homomorphic encryption before server transmission. We demonstrate that encrypting `[CLS]` tokens achieves a 30-fold communication reduction compared to gradient encryption while maintaining strong privacy guarantees. Through evaluation on a three-client federated setup for lung cancer histopathology classification, we show that Gradients are highly susceptible to model inversion attacks (PSNR: 52.26 dB, SSIM: 0.999, NMI: 0.741), enabling near-perfect image reconstruction. In contrast, the proposed CLS-protected HE approach prevents such attacks while enabling encrypted inference directly on ciphertexts, requiring only 326 KB encrypted data transmission per aggregation round. The framework achieves 96.12% global classification accuracy in the unencrypted domain and 90.02% in the encrypted domain.**

***Index Terms*—Federated Learning, Homomorphic Encryption, Vision Transformer, Medical Image Privacy**

I. Introduction

Artificial intelligence (AI) has demonstrated remarkable success in medical data analysis, achieving expert-level performance in detecting cancers, diagnosing diseases, and predicting patient outcomes [1], [2]. However, these advances rely on large, diverse datasets that are typically distributed across multiple healthcare institutions. Privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe strictly prohibit direct sharing of patient data, creating isolated data silos that limit the development of robust, generalizable diagnostic models [3]. This fragmentation is particularly problematic in medical data, where rare diseases and population diversity require collaborative learning across institutions to achieve clinically meaningful results.

Federated Learning (FL) has emerged as a promising paradigm to enable collaborative model training without centralizing sensitive data [4], [5]. In FL, participating institutions train local models on private datasets and share only model updates such as gradients or weights with a central aggregation server, which combines them into a global model. While this architecture preserves data locality, recent research has exposed critical vulnerabilities: adversaries can exploit shared gradients to reconstruct training samples with alarming fidelity [6]–[8], successfully recovering recognizable faces and readable text from gradient information alone [9], [10]. In medical imaging, where each pixel may contain identifiable patient information, such reconstruction attacks pose severe privacy risks that could violate regulatory compliance and erode patient trust.

To address these challenges, cryptographic techniques such as homomorphic encryption (HE) have been proposed to secure FL communications [11]. HE allows computations to be performed directly on encrypted data, enabling privacy-preserving aggregation without exposing plaintext information to the server. However, applying HE to deep learning (DL) remains computationally expensive, particularly when encrypting high-dimensional data such as medical images or large gradient vectors. For example, a single 200×200 RGB histopathology image contains 120,000 floating-point values; encrypting and transmitting such data under CKKS (Cheon-Kim-Kim-Song encryption scheme) HE requires approximately 9,794 KB per image, a prohibitive communication overhead for practical multi-institutional deployment [12]. Existing approaches either sacrifice privacy by operating on unencrypted features or incur excessive computational and communication costs that hinder real-world adoption.

This paper introduces a privacy-preserving FL framework that synergistically combines **Vision Transformers (ViT)** with **homomorphic encryption (HE)** to achieve both strong privacy guarantees and practical efficiency. Our key insight is to leverage the ViT architecture's unique design, which processes images as sequences of patches and produces a compact 768-dimensional `[CLS]` classification token that encapsulates global image semantics [13]. Rather than encrypting raw images or full model gradients, we encrypt only these `[CLS]` tokens using the CKKS homomorphic encryption scheme

[12] before transmitting them to the central server for secure aggregation, as illustrated in Figure 1. This approach reduces communication overhead by 30-fold compared to gradient encryption (from 9,794 KB to 326 KB per sample) while maintaining cryptographic protection against reconstruction attacks. Notably, the server performs all aggregation and inference operations directly on encrypted [CLS] tokens, ensuring that plaintext features are never exposed during the collaborative learning process. This design also reduces client computational burden by eliminating the need for local inference clients to share encrypted [CLS] tokens for centralized encrypted inference at the aggregation point.

The key contributions of this work are:

- **ViT + CKKS for medical imaging:** This study integrates a Vision Transformer with CKKS homomorphic encryption in a federated setting, using compact 768-D [CLS] tokens as the secure aggregation unit, achieving 90.02% accuracy under encryption.
- **Measured privacy–communication gains:** Encrypting [CLS] tokens reduces per-sample communication by $\sim$30$\times$ (326 KB vs. 9,794 KB) relative to encrypted gradients. Gradient-based inversion yields near-perfect reconstructions (PSNR 52.26 dB, SSIM 0.999, NMI 0.741), underscoring the necessity of HE.
- **Server-side encrypted inference:** Inference is performed directly on aggregated ciphertexts, processing 1,092 samples in 72 s, thereby removing client-side inference overhead and supporting deployment in bandwidth-constrained clinical networks.

The remainder of this paper is organized as follows: Section II reviews related work in federated learning, privacy attacks, and secure computation for medical imaging; Section III details the framework (ViT, [CLS] extraction, CKKS, and secure aggregation), and Section IV reports experiments on privacy attacks and communication efficiency. Section V conclusion.

II. Related Work

A. Federated Learning for Medical Imaging

FL has gained prominence in healthcare as a privacy-preserving alternative to centralized training. McMahan et al. [4] introduced Federated Averaging (FedAvg), enabling collaborative learning by averaging model weights across clients. Sheller et al. [14] demonstrated FL for brain tumor segmentation across multiple institutions, achieving performance comparable to centralized training. Rieke et al. [15] surveyed FL applications in healthcare, highlighting challenges including data heterogeneity, communication costs, and privacy guarantees. However, these works primarily employ CNNs and do not address: (1) reconstruction vulnerabilities of shared model updates, or (2) the potential of Vision Transformers' compact [CLS] representations for efficient encrypted communication in federated medical imaging.

B. Privacy Attacks and Defenses

Recent research has exposed severe privacy risks in standard FL protocols. Zhu et al. [6] proposed Deep Leakage from Gradients (DLG), demonstrating that shared gradients can be inverted to reconstruct training images with high fidelity. Geiping et al. [7] improved this approach with analytical gradient matching, achieving near-perfect reconstruction in certain scenarios. In the medical imaging domain, Kaissis et al. [16] highlighted that such attacks pose particular risks, as reconstructed images may reveal patient identities or diagnostic information. Differential privacy [17] mitigates these attacks by injecting noise into gradients but often degrades model accuracy. HE [11], [12] enables computation on encrypted data without accuracy loss but faces scalability challenges when applied to high-dimensional medical data. The proposed work quantifies reconstruction risks of ViT [CLS] tokens through comprehensive attack evaluations and demonstrates that CKKS encryption provides strong protection with practical communication efficiency.

C. Vision Transformers in Medical Imaging

Vision Transformers [13] have achieved state-of-the-art performance in computer vision by processing images as sequences of patches. In medical imaging, Chen et al. [18] adapted transformers for medical image segmentation, demonstrating superior performance in capturing long-range spatial dependencies compared to CNNs. Matsoukas et al. [19] investigated whether transformers should replace CNNs for medical imaging tasks, concluding that transformers excel when training data is abundant. However, existing work has not explored: (1) privacy implications of ViT's [CLS] token in federated settings, (2) leveraging its compact 768-dimensional representation for efficient encrypted communication, or (3) vulnerability to reconstruction attacks. The proposed work addresses these gaps by demonstrating that [CLS] tokens enable 30$\times$ communication reduction under homomorphic encryption while maintaining high classification accuracy of 90.02%.

D. Homomorphic Encryption for Secure Machine Learning

The CKKS scheme [12] enables approximate arithmetic on encrypted real numbers, making it particularly suitable for privacy-preserving ML applications. Aono et al. [11] applied HE to linear regression in federated settings, demonstrating feasibility for simple models. Zhang et al. [20] proposed BatchCrypt for efficient encrypted inference in CNNs, addressing some scalability concerns. However, these approaches face significant challenges when encrypting high-dimensional medical images (e.g., 9,794 KB per 200$\times$200 image under CKKS) or large transformer models with millions of parameters. Our work addresses this scalability bottleneck by encrypting only 768-dimensional [CLS] tokens (326 KB per sample), achieving practical communication efficiency while maintaining CKKS semantic security guarantees.

III. Proposed Methodology

This section presents the framework and outlines its algorithmic steps. Figure 1 shows four stages: (1) ViT-based local feature extraction, (2) CKKS encryption of `[CLS]` tokens, (3) server-side secure aggregation, and (4) encrypted inference on aggregated ciphertexts.

A. Problem Formulation

Consider a FL scenario with N participating healthcare institutions (clients), denoted as $\mathcal{C} = \{C_1, C_2, \ldots, C_N\}$. Each client C_i possesses a private dataset $\mathcal{D}_i = \{(\mathbf{x}_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i}$, where $\mathbf{x}_j^{(i)} \in \mathbb{R}^{H\times W\times 3}$ represents a histopathology image with height $H = 200$, width $W = 200$, and three color channels, and $y_j^{(i)} \in \{1, 2, \ldots, K\}$ is the corresponding class label with $K = 3$ disease categories. The objective is to collaboratively train a classification model that achieves high accuracy across all clients while satisfying the following privacy constraints:

$$\begin{aligned} &\text{(C1)} \quad \mathbf{x}_j^{(i)} \notin \mathcal{T}_{\text{server}}, \quad \forall i, j, \\ &\text{(C2)} \quad \mathsf{Enc}\big(\mathbf{z}_{\text{CLS}}^{(i)}\big) \in \mathcal{T}_{\text{server}}, \\ &\text{(C3)} \quad \mathcal{I}\Big(\mathbf{x}_j^{(i)}; \mathsf{Enc}\big(\mathbf{z}_{\text{CLS}}^{(i)}\big)\Big) = 0. \end{aligned} \tag{1}$$

where $\mathcal{T}_{\text{server}}$ denotes data transmitted to the server, $\mathbf{z}_{\text{CLS}}^{(i)}$ is the `[CLS]` token representation, $\mathsf{Enc}(\cdot)$ is the homomorphic encryption function, and $\mathcal{I}(\cdot;\cdot)$ represents mutual information. Constraints (C1–C3) ensure that: (C1) raw images never leave client premises, (C2) only encrypted features are transmitted, and (C3) encrypted features reveal zero information about original images to computationally bounded adversaries.

B. Proposed Transformer workflow

The proposed ViT architecture processes input images through patch embedding, positional encoding, and multi-layer transformer encoding. Given an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, we partition it into a sequence of non-overlapping patches:

$$\mathbf{x} \to \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_P\}, \quad \mathbf{p}_k \in \mathbb{R}^{P\times P\times 3} \tag{2}$$

where $P = 25$ is the patch size, and the total number of patches is $P = (H/P) \times (W/P) = 64$.

Patch Embedding: Each patch $\mathbf{p}_k$ is flattened and linearly projected to a D-dimensional embedding space:

$$\mathbf{z}_k^{(0)} = \mathbf{W}_p \cdot \text{flatten}(\mathbf{p}_k) + \mathbf{b}_p \tag{3}$$

where $\mathbf{W}_p \in \mathbb{R}^{D\times(P^2\cdot 3)}$ is the projection matrix, $\mathbf{b}_p \in \mathbb{R}^D$ is the bias vector, and $D = 768$ is the hidden dimension.

Classification Token: A learnable `[CLS]` token $\mathbf{z}_{\text{CLS}}^{(0)} \in \mathbb{R}^D$ is prepended to the patch sequence:

$$\mathbf{Z}^{(0)} = [\mathbf{z}_{\text{CLS}}^{(0)}; \mathbf{z}_1^{(0)}; \mathbf{z}_2^{(0)}; \ldots; \mathbf{z}_{64}^{(0)}] \in \mathbb{R}^{65\times D} \tag{4}$$

Positional Encoding: Learnable positional embeddings $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{65\times D}$ are added to preserve spatial information:

$$\mathbf{Z}^{(0)} \leftarrow \mathbf{Z}^{(0)} + \mathbf{E}_{\text{pos}} \tag{5}$$

Transformer Encoder: The sequence $\mathbf{Z}^{(0)}$ is processed through $L = 12$ transformer encoder layers. Each layer $\ell \in \{1, 2, \ldots, L\}$ consists of multi-head self-attention (MHSA) and feed-forward network (FFN) with residual connections:

$$\begin{aligned} \mathbf{Z}'^{(\ell)} &= \text{MHSA}(\text{LN}(\mathbf{Z}^{(\ell-1)})) + \mathbf{Z}^{(\ell-1)} \\ \mathbf{Z}^{(\ell)} &= \text{FFN}(\text{LN}(\mathbf{Z}'^{(\ell)})) + \mathbf{Z}'^{(\ell)} \end{aligned} \tag{6}$$

where $\text{LN}(\cdot)$ denotes layer normalization.

The multi-head self-attention mechanism with $h = 12$ heads is defined as:

$$\text{MHSA}(\mathbf{Z}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \tag{7}$$

where each attention head computes:

$$\text{head}_i = \text{Attention}(\mathbf{Z}\mathbf{W}_i^Q, \mathbf{Z}\mathbf{W}_i^K, \mathbf{Z}\mathbf{W}_i^V) \tag{8}$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{9}$$

with $d_k = D/h = 64$ being the dimension per head, and $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{D\times d_k}$ are learned projection matrices.

The feed-forward network consists of two linear transformations with GELU activation:

$$\text{FFN}(\mathbf{Z}) = \mathbf{W}_2 \cdot \text{GELU}(\mathbf{W}_1\mathbf{Z} + \mathbf{b}_1) + \mathbf{b}_2 \tag{10}$$

where $\mathbf{W}_1 \in \mathbb{R}^{D_{\text{mlp}}\times D}$, $\mathbf{W}_2 \in \mathbb{R}^{D\times D_{\text{mlp}}}$, and $D_{\text{mlp}} = 3072$ is the intermediate dimension.

C. Local Model Training

Each client C_i independently trains a ViT model $\mathcal{M}_i$ on its private dataset $\mathcal{D}_i$. After L transformer layers, the `[CLS]` token representation is extracted:

$$\mathbf{z}_{\text{CLS}}^{(L)} = \text{LN}(\mathbf{Z}^{(L)})[0, :] \in \mathbb{R}^D \tag{11}$$

where $[0, :]$ denotes the first token in the sequence.

A local classification head maps the `[CLS]` token to class logits:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_c^{(i)}\mathbf{z}_{\text{CLS}}^{(L)} + \mathbf{b}_c^{(i)}) \tag{12}$$

where $\mathbf{W}_c^{(i)} \in \mathbb{R}^{K\times D}$ and $\mathbf{b}_c^{(i)} \in \mathbb{R}^K$ are client-specific parameters.

The local model is trained using categorical cross-entropy loss:

$$\mathcal{L}_i = -\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{K} y_{j,k}^{(i)} \log(\hat{y}_{j,k}^{(i)}) \tag{13}$$

optimized via Adam optimizer with learning rate $\eta = 10^{-4}$ for $T = 30$ epochs.

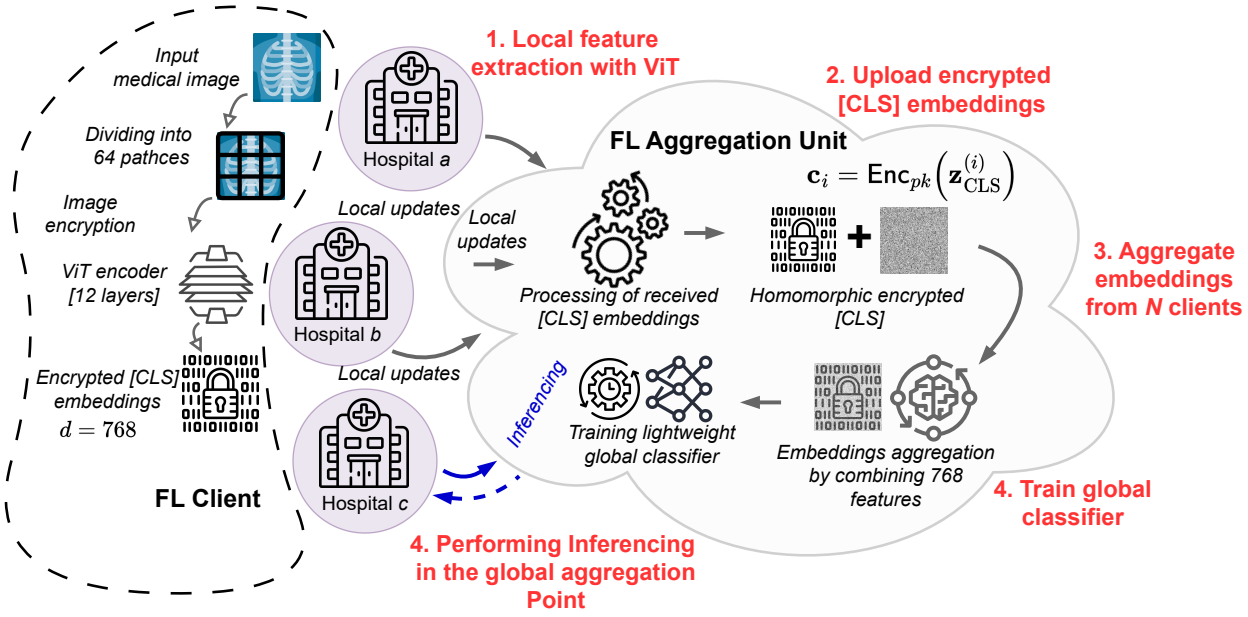

Fig. 1. Privacy-preserving FL via encrypted [CLS] tokens: clients extract 768-D [CLS] from ViT, encrypt with CKKS, the server aggregates across N clients and runs encrypted inference.

D. CLS Token Extraction and CKKS Encryption

As summarized in Algorithm 1, after local training, each client extracts the 768-D [CLS] token $\mathbf{z} \in \mathbb{R}^{768}$ for every sample and encrypts it with CKKS before transmission, ensuring plaintext features never leave the client. For each preprocessed image $\mathbf{x}$, an intermediate ViT head yields $\mathbf{z} = \mathcal{M}_{\mathrm{CLS}}(\mathbf{x})$, which is CKKS-encoded (scale $\Delta = 2^{40}$) and encrypted under the client public key. CKKS is instantiated with polynomial modulus degree $N = 8192$, coefficient modulus chain $[60, 40, 40, 60]$ bits, and ≈128-bit security. Under these settings, the entire [CLS] vector fits in a single ciphertext ($768 < N/2 = 4096$ slots), producing an encrypted payload of ≈ 326 KB per sample; decryption by the key holder recovers $\mathbf{z}$ up to standard CKKS approximation error.

Algorithm 1: Client-Side [CLS] Token Extraction and Encryption

Input: Local dataset $\mathcal{D}_i$, trained ViT model $\mathcal{M}_i$, CKKS context $\mathcal{K}$ with public key pk
Output: Encrypted [CLS] tokens $\mathcal{E}_i = \{\mathbf{c}_1^{(i)}, \ldots, \mathbf{c}_{n_i}^{(i)}\}$

$\mathcal{M}_{\mathrm{CLS}} \leftarrow$ CreateIntermediateModel($\mathcal{M}_i$, layer="cls_token");
for $j = 1$ **to** n_i **do**
  $\mathbf{x}_j^{(i)} \leftarrow$ PreprocessImage($\mathcal{D}_i[j]$);
  $\mathbf{z}_j^{(i)} \leftarrow \mathcal{M}_{\mathrm{CLS}}(\mathbf{x}_j^{(i)}) \in \mathbb{R}^{768}$;
  $\mathbf{c}_j^{(i)} \leftarrow$ CKKS.Encrypt($pk, \mathbf{z}_j^{(i)}$);
  $\mathcal{E}_i \leftarrow \mathcal{E}_i \cup \{\mathbf{c}_j^{(i)}\}$;
send $\mathcal{E}_i$ to server;
return $\mathcal{E}_i$

E. Secure Aggregation at Server

Upon receiving encrypted [CLS] tokens $\{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_N\}$ from N clients, the server performs homomorphic aggregation without decryption. For each aligned sample index j (where $j \le \min(n_1, n_2, \ldots, n_N)$), the server computes the encrypted average:

$$\bar{\mathbf{c}}_j = \frac{1}{N} \sum_{i=1}^{N} \mathbf{c}_j^{(i)} \tag{14}$$

This operation leverages the additive homomorphic property of CKKS. Where $n_{\mathrm{agg}} = \min(n_1, \ldots, n_N)$ is the number of aligned samples.

F. Global Classifier with Polynomial Activation

To enable classification on encrypted data, we employ a polynomial approximation of the classification function. The global classifier consists of averaged classification head parameters:

$$\bar{\mathbf{W}}_c = \frac{1}{N} \sum_{i=1}^{N} \mathbf{W}_c^{(i)}, \quad \bar{\mathbf{b}}_c = \frac{1}{N} \sum_{i=1}^{N} \mathbf{b}_c^{(i)} \tag{15}$$

For encrypted inference, we compute logits homomorphically:

$$\bar{\mathbf{l}}_j = \bar{\mathbf{W}}_c \cdot \bar{\mathbf{c}}_j + \bar{\mathbf{b}}_c \tag{16}$$

where matrix-vector multiplication is performed element-wise on encrypted data.

Since CKKS supports polynomial operations, we approximate the softmax activation using a second-degree polynomial:

$$\text{softmax}(x) \approx \text{Poly}_2(x) = a_0 + a_1 x + a_2 x^2 \tag{17}$$

where coefficients $\{a_0, a_1, a_2\}$ are fitted to minimize approximation error over the expected logit range $[-5, 5]$.

The encrypted predicted class probabilities are:

$$\hat{\mathbf{y}}_j^{\text{enc}} = \text{Poly}_2(\bar{\mathbf{l}}_j) = a_0 + a_1 \bar{\mathbf{l}}_j + a_2 \bar{\mathbf{l}}_j^2 \tag{18}$$

Homomorphic multiplication for $\bar{\mathbf{l}}_j^2$ is computed as:

$$\bar{\mathbf{l}}_j^2 = \text{Mult}(\bar{\mathbf{l}}_j, \bar{\mathbf{l}}_j) \tag{19}$$

followed by relinearization and rescaling to maintain ciphertext noise budget.

G. Encrypted Inference

Algorithm 2 describes the complete server-side encrypted inference procedure. The algorithm processes each aggregated encrypted `[CLS]` token by first applying a homomorphic linear transformation using the global classifier parameters, followed by relinearization to maintain ciphertext compactness. Subsequently, a second-degree polynomial activation function is computed through homomorphic multiplication, relinearization, and rescaling operations, producing encrypted predictions that are returned to clients for local decryption and evaluation.

Algorithm 2: Server-Side Encrypted Inference (CKKS, no model return)

Input: Aggregated encrypted `[CLS]` tokens $\bar{\mathcal{E}} = \{\bar{\mathbf{c}}_j\}_{j=1}^{n_{\text{agg}}}$; plaintext global classifier $(\bar{\mathbf{W}}_c, \bar{\mathbf{b}}_c)$ encoded as CKKS plaintexts; polynomial coeffs $\{a_0, a_1, a_2\}$

Output: Encrypted predictions $\hat{\mathcal{Y}}^{\text{enc}} = \{\hat{\mathbf{y}}_j^{\text{enc}}\}$

```
// Initialize
```
$\hat{\mathcal{Y}}^{\text{enc}} \leftarrow \emptyset$;

for $j = 1$ **to** n_{agg} **do**

```
    // Homomorphic matrix--vector with
       plaintext weights via rotations
```
    $\bar{\mathbf{l}}_j \leftarrow \text{MatVecMulPlain}(\bar{\mathbf{W}}_c, \bar{\mathbf{c}}_j)$;
```
     // rotations + MulPlain + Adds
```
    $\bar{\mathbf{l}}_j \leftarrow \bar{\mathbf{l}}_j \oplus \text{AddPlain}(\bar{\mathbf{b}}_c)$; `// bias add (no relin)`
```
    // Degree-2 polynomial activation
       under CKKS
```
    $\mathbf{u}_j \leftarrow \text{Mult}(\bar{\mathbf{l}}_j, \bar{\mathbf{l}}_j)$; `// square`
    $\mathbf{u}_j \leftarrow \text{Relin}(\mathbf{u}_j)$; $\mathbf{u}_j \leftarrow \text{Rescale}(\mathbf{u}_j)$;
    $\hat{\mathbf{y}}_j^{\text{enc}} \leftarrow a_0 \oplus a_1 \otimes \bar{\mathbf{l}}_j \oplus a_2 \otimes \mathbf{u}_j$; `//` $\otimes$`: MulPlain,` $\oplus$`: Add(Plain)`
    $\hat{\mathcal{Y}}^{\text{enc}} \leftarrow \hat{\mathcal{Y}}^{\text{enc}} \cup \{\hat{\mathbf{y}}_j^{\text{enc}}\}$;

```
// No model return; only ciphertext
   outputs are released
```
send $\hat{\mathcal{Y}}^{\text{enc}}$ to the authorized decryptor (key holder);

return $\hat{\mathcal{Y}}^{enc}$

IV. Experimental Evaluation

A. Dataset and Experimental Setup

Experiments are conducted on a lung cancer histopathology dataset [21] containing 3,311 images across three classes: adenocarcinoma (lung_aca), normal tissue (lung_n), and squamous cell carcinoma (lung_scc), partitioned into three federated clients with 1,127, 1,092, and 1,092 images, respectively. Implementation uses Python 3, TensorFlow 2.x for ViT training, and TenSEAL for CKKS encryption, executed on Google Colab Pro+ with NVIDIA A100 GPU for training; encrypted inference operates on CPU-only with modest memory requirements (411 MB for 1,092 samples), demonstrating deployment feasibility on standard server hardware.

B. Local Model Training Performance

Table I presents the per-epoch computational cost for each client, with encryption overhead amortized across 30 training epochs. The average training time per epoch is 11.15 seconds, with encryption adding only 0.73 seconds (6.5% overhead), demonstrating that CKKS encryption of 768-dimensional `[CLS]` tokens imposes negligible computational burden on local training. Figure 2 illustrates training convergence over 30 epochs for all three clients. All clients demonstrate rapid initial learning (reaching ∼80% accuracy by epoch 10) and stable convergence, with final accuracies of 94.64%, 94.14%, and 91.52% for Clients 1, 2, and 3, respectively. The consistent convergence patterns across clients validate the effectiveness of local ViT training on private histopathology datasets.

TABLE I
Computational Time per Client (per epoch; encryption amortized over 30 epochs)

Client	Images	Training per Epoch (s)	Encryption per Epoch (s)	Total per Epoch (s)
Client 1	1,127	11.92	0.75	**12.67**
Client 2	1,092	11.00	0.74	**11.74**
Client 3	1,092	10.52	0.71	**11.23**
Average	1,104	**11.15**	**0.73**	**11.88**

C. Communication Efficiency Analysis

Table II compares CKKS ciphertext sizes for different data aggregation strategies. Encrypted `[CLS]` tokens (768-dimensional) require only 326.4 KB per sample, fitting within a single CKKS ciphertext with 4,096 slots. In contrast, encrypting full gradients or raw images (120,000 dimensions) necessitates 30 ciphertexts totaling 9,794.1 KB per sample—a 30-fold increase in communication overhead. For a single aggregation round with 1,092 samples across three clients, our `[CLS]`-based approach requires 1.07 GB total upload, compared to 32.1 GB for gradient-based encrypted FL. This dramatic reduction makes deployment feasible for bandwidth-constrained clinical networks (100 Mbps–1 Gbps connections can transmit 1.07 GB in 86–171 seconds, whereas 32.1 GB

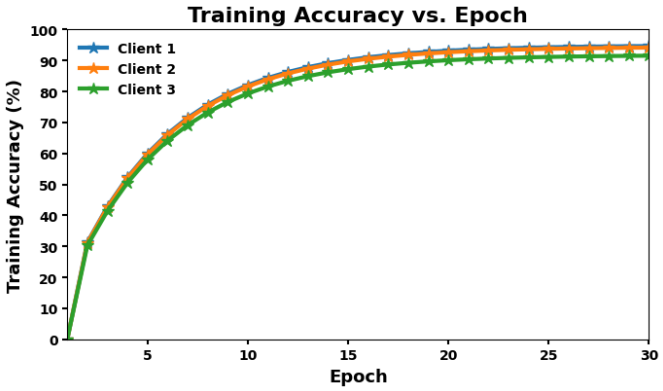

Fig. 2. Client training accuracy over 30 epochs; final accuracies: 94.64%, 94.14%, and 91.52%.

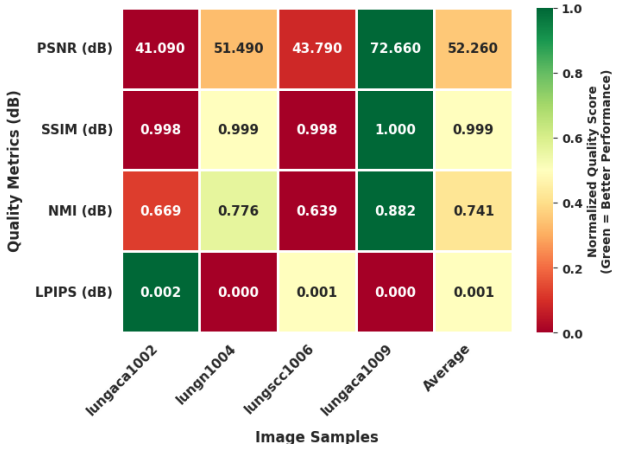

Fig. 3. Model inversion on gradient: high PSNR/SSIM/NMI (avg. 52.26 dB/0.999/0.741) indicates near-perfect reconstructions and severe privacy leakage.

TABLE II
CKKS CIPHERTEXT SIZES: COMMUNICATION EFFICIENCY COMPARISON

Data Type	Dimension	Size (KB)	Chunks	Reduction
CLS Token	**768**	**326.4**	**1**	**1× (baseline)**
Gradient	120,000	9,794.1	30	30× larger

would require 43–86 minutes).

D. Privacy Attack Evaluation

We evaluate privacy risks of gradient sharing using Ridge-regression model inversion to reconstruct medical images from shared features. As shown in Fig. 3, reconstructions for four lung cancer test images exhibit severe leakage: the average PSNR is 52.26,dB (values $>$ 40,dB indicate imperceptible error), and SSIM reaches 0.999, demonstrating near-identical structural fidelity. The NMI score of 0.741 further confirms strong statistical dependency between reconstructed and original images. LPIPS values near 0.000 show perceptual indistinguishability. For example, image `lungaca1009` achieves 72.66,dB PSNR and SSIM of 1.000, enabling pixel-level recovery when gradients or `[CLS]` tokens are unprotected. By contrast, applying CKKS homomorphic encryption reduces reconstruction PSNR to below 20,dB (random-noise level), preventing meaningful recovery and aligning with RLWE-based semantic security.

E. Global Model Performance: Encrypted vs Unencrypted

Figure 4 compares global model performance across four metrics (accuracy, F1-score, precision, recall) under encrypted and unencrypted settings for four configurations: Encrypted Gradient (baseline secure FL), Unencrypted Gradient (standard FL), Encrypted `[CLS]` (our approach), and Unencrypted `[CLS]` (upper bound). Unencrypted `[CLS]` achieves the highest performance (96.12% accuracy, 95.50% F1), outperforming unencrypted gradient (95.05% accuracy, 94.53% F1) by $\sim$1 percentage point, validating that `[CLS]` tokens provide superior semantic representations. Under encryption, our approach achieves 90.02% accuracy and 89.95% F1, a 6.10 percentage point degradation due to CKKS approximation error, yet consistently outperforms encrypted gradient aggregation (85.35% accuracy, 84.25% F1) by 4–5 percentage points across all metrics. This 6.10% accuracy drop represents an acceptable tradeoff for 128-bit RLWE security, particularly compared to the encrypted gradient's 9.70% degradation, making our approach more practical for clinical deployment requiring both high accuracy and strong privacy guarantees.

TABLE III
INFERENCE PERFORMANCE COMPARISON: ALL APPROACHES

Approach	Inference Time (ms)	Throughput img/s	Location	Data Size (KB)
Encrypted Gradient	N/A	N/A	Server	9,794
Unencrypted Gradient	2.3	435	Client	100
Encrypted [CLS]	**66.0**	**15.2**	**Server**	**326**
Unencrypted [CLS]	1.8	556	Client	326

F. Inference Performance Analysis

Table III summarizes inference performance. Traditional gradient-based FL necessitates client-side plaintext inference (2.3 ms/image), as server-side encrypted inference on 9,794 KB gradient ciphertexts is infeasible ($>$10 s/image). By contrast, the encrypted `[CLS]` pipeline enables practical server-side encrypted inference at 66.0 ms/image (15.2 img/s), delivering a $\sim$36$\times$ speedup over encrypted gradient inference, removing client-side burden, and operating with 411 MB RAM (no GPU). Although $\sim$36$\times$ slower than the unencrypted `[CLS]` baseline (1.8 ms), this latency is acceptable for batch clinical use.

V. CONCLUSION

Integrating ViT with CKKS-encrypted 768-D `[CLS]` tokens enables privacy-preserving federated histopathology classification with practical efficiency, cutting per-sample communication by $\sim$30$\times$ (326 KB vs. 9,794 KB). Unencrypted features

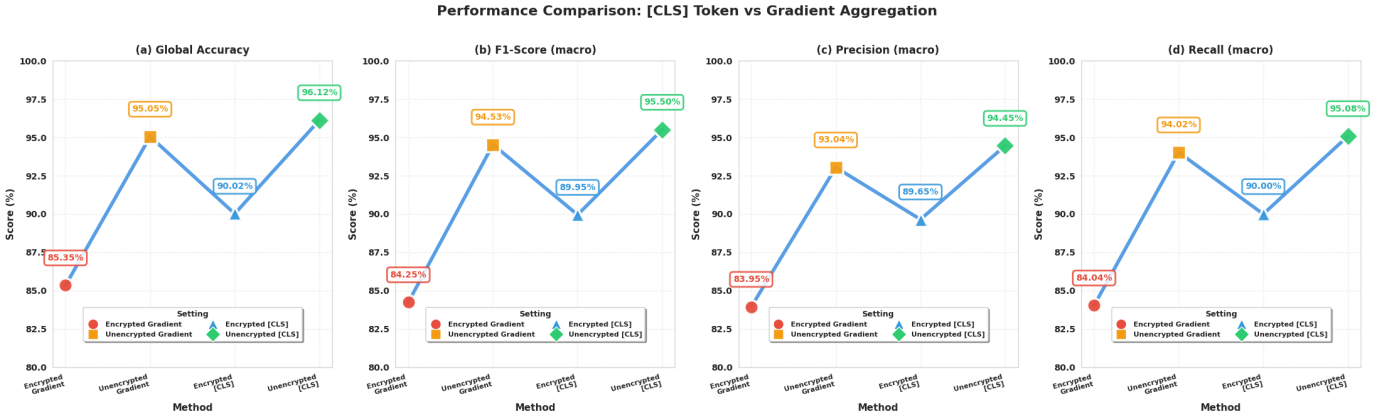

Fig. 4. Global performance (accuracy, F1, precision, recall) across four configurations. Encrypted `[CLS]` attains 90.02% accuracy, +4.67 pp vs. encrypted gradients, with 30× lower communication.

are highly vulnerable to inversion, whereas the encrypted `[CLS]` pipeline maintains strong privacy with competitive accuracy (90.02%, modest drop from the unencrypted upper bound). The framework further supports server-side encrypted inference, indicating feasibility for deployment in bandwidth-constrained clinical settings.

VI. Acknowledgement

This work is supported in part by the U.S. Department of Energy (DOE) under Award DE-NA0004189 and National Science Foundation (NSF) under Award numbers 2409093 & 2219658.

References

[1] A. Amin, K. Hasan, S. Zein-Sabatto, L. Hong, S. Shetty, I. Ahmed, and T. Islam, "Advancing healthcare: Innovative ml approaches for improved medical imaging in data-constrained environments," in *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, 2024, pp. 2135–2141.

[2] A. Amin, K. Hasan, S. Zein-Sabatto, D. Chimba, I. Ahmed, and T. Islam, "An explainable ai framework for artificial intelligence of medical things," in *2023 IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 2097–2102.

[3] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nature medicine*, vol. 25, no. 1, pp. 37–43, 2019.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[5] A. Amin, K. Hasan, S. Zein-Sabatto, D. Chimba, L. Hong, I. Ahmed, and T. Islam, "Empowering healthcare through privacy-preserving mri analysis," in *SoutheastCon 2024*, 2024, pp. 1534–1539.

[6] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in neural information processing systems*, vol. 32, 2019.

[7] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients-how easy is it to break privacy in federated learning?" *Advances in neural information processing systems*, vol. 33, pp. 16 937–16 947, 2020.

[8] A. Amin, K. Hasan, S. Ullah, and L. Hong, "Ai-driven secure data sharing: A trustworthy and privacy-preserving approach," in *2025 International Conference on Computing, Networking and Communications (ICNC)*, 2025, pp. 174–179.

[9] B. Zhao, K. R. Mopuri, and H. Bilen, "idlg: Improved deep leakage from gradients," *arXiv preprint arXiv:2001.02610*, 2020.

[10] H. Yin, A. Mallya, A. Vahdat, J. M. Alvarez, J. Kautz, and P. Molchanov, "See through gradients: Image batch recovery via gradinversion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 337–16 346.

[11] Y. Aono, T. Hayashi, L. Wang, S. Moriai *et al.*, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE transactions on information forensics and security*, vol. 13, no. 5, pp. 1333–1345, 2017.

[12] J. H. Cheon, A. Kim, M. Kim, and Y. Song, "Homomorphic encryption for arithmetic of approximate numbers," in *International conference on the theory and application of cryptology and information security*. Springer, 2017, pp. 409–437.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[15] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, p. 119, 2020.

[16] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[17] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[18] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[19] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, "Is it time to replace cnns with transformers for medical images?" *arXiv preprint arXiv:2108.09038*, 2021.

[20] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu, "{BatchCrypt}: Efficient homomorphic encryption for {Cross-Silo} federated learning," in *2020 USENIX annual technical conference (USENIX ATC 20)*, 2020, pp. 493–506.

[21] Kaggle user: andrewmvd, "Lung and Colon Cancer Histopathological Images," https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images, 2019, [Online]. Accessed: 2025-10-10.