Federated Learning with Adaptive Gradient Compression and Dynamic Aggregation for Privacy-Preserving Small-Sample Data

Leiqian Qi

Xuzhou University of Technology, Xuzhou, Jiangsu Province, 221000, China
E-mail: leiqianqi632@gmail.com

Received 20 October 2025; Accepted 14 April 2026

Abstract

This paper proposes a federated learning (FL) framework that incorporates adaptive gradient compression and dynamic aggregation to address communication efficiency and data privacy issues in the context of FL with limited sample size and non-IID data distributions in edge devices and resource-scarce environments. This proposed framework incorporates dynamic gradient compression techniques that compress gradients based on their magnitude and variance to ensure high communication efficiency with minimal loss in model accuracy. Meanwhile, the proposed framework incorporates dynamic aggregation techniques that assign different weights to clients based on their reliability to ensure effective model convergence in heterogeneous and scarce data distributions. Data privacy in the proposed framework is ensured through secure aggregation and Differential Privacy (DP) techniques. Experimental results on various datasets, including LEAF, FEMNIST, Reddit, and Shakespeare, show that the proposed framework ensures communication efficiency of over 70%, preserves model accuracy with minimal loss at 1–2%, and achieves 30% faster convergence speed

Journal of Cyber Security and Mobility, Vol. 15_3, 709–746.
doi: 10.13052/jcsm2245-1439.1538

compared to traditional FL techniques. These results show that the proposed framework is applicable in real-world scenarios in mobile edge computing and IoT applications, where communication efficiency and data privacy are significant factors for model convergence and deployment. The combination of gradient compression and dynamic aggregation in FL with strong privacy guarantees makes this framework a powerful tool for model convergence in heterogeneous scenarios.

Keywords: Federated learning, privacy protection, differential privacy, secure aggregation, gradient compression, dynamic aggregation.

1 Introduction

The exponential growth of data and the increasing concern for user privacy have intensified the demand for distributed learning methods that do not require centralized data storage [1]. Federated learning (FL) effectively addresses this need via a collaborative training model, in which the local data is not exposed [2]. This method utilizes adaptive gradient compression to compress the gradients based on their size. It also utilizes dynamic aggregation to weight the clients' updates based on their reliability. All of these methods are combined to form the FL framework. However, FL deals with serious challenges in small-sample data. When limited data is available, FL witnesses convergence and generalization issues. These issues are significant within the resource-constrained environments of the device communication and central servers [3]. These challenges are addressed by integrating the adaptive gradient compression and aggregation strategies. This integration also reduces the communication cost and improves learning efficiency [4]. Subsequently, the integrated form of FL, adaptive gradient compression, and dynamic aggregation offers enhanced privacy protection and performance in the case of a small data source. Another response to the small data source is the symbolic AI and knowledge approaches. By incorporating expert-defined rules, logic-based inference, and ontologies methods, the lack of sufficient data during the training phase is addressed [5]. In these methods, decisions are made through the domain-specific rules to elevate the explainability and interpretability [6]. Also, these methods provide a structured framework for integrating prior knowledge into learning models via the semantic networks and logic trees characteristics [7]. Despite these advantages, symbolic methods have adaptability and scalability issues in the noisy, highdimensional, and continuously evolving data. Therefore, a

data-driven technique with more flexibility is required to assess diverse and dynamic scenarios [8]. In this regard, statistical inference and data-driven optimization have emerged [9]. The decision trees, support vector machines, and ensemble methods offer an enhanced performance within moderate-sized datasets [10]. These methods enable local model training and aggregation without the need for explicit rule encoding. However, issues of overfitting and lack of generalizability emerge due to insufficient data for training [11]. Since these methods depend on handcrafted features, their ability to extract rich representations from raw data is limited. Besides, the aggregation algorithm of FedAvg is introduced to combine models, while data heterogeneity and the varying contribution of individuals are eliminated [12]. Deep learning and pretrained models are developed to address the generalization and feature representation issues of traditional machine learning methods [13]. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based architectures offer remarkable capabilities in learning from high-dimensional data and transferring knowledge across tasks [14]. These models are adapted to handle decentralized training through techniques in FL. Also, communication overhead is reduced via compression techniques, including quantization, smartification, and adaptive gradient clipping. However, these methods also suffer from performance degradation within the small and non-IID data sources [15]. Besides, the standard aggregation mechanisms are not able to respond to client variability and optimize bandwidth utilization. Subsequently, the need for more adaptive models with hierarchical representation and dynamic adjustment to non-IID data distributions is highlighted.

The current investigation addresses the limitations of fixed aggregation strategies and high communication overhead in a deep FL. An integrated adaptive gradient compression with a dynamic aggregation algorithm is proposed to enhance communication efficiency and model performance while preserving data privacy in small and heterogeneous datasets. Utilizing adaptive gradient compression has a great impact on cutting down the transmission load. The dynamic aggregation algorithm gives a higher weight to client data according to the contributions of quality and gradients to the informative updates; thus, effective informative updates are ensured. This twin mechanism is no longer the most effective complement to training stability but additionally leads to faster convergence of the model and will increase its generalization across distinct clients. Therefore, the proposed approach gives a technique for the essential compromise among privacy, accuracy, and performance in an FL setting.

This paper introduces an adaptive gradient compression approach, adjusting the compression tiers dynamically in step with the local records variance and communication bandwidth. A dynamic aggregation strategy is recommended for giving special weights to purchaser updates, to make the system more robust against information heterogeneity, and concurrently will lead to better version convergence. Results imply that accuracy has multiplied and communication performance has improved as compared to conventional FL procedures. These findings are based on many restricted-pattern benchmark datasets.

2 Related Work

2.1 Adaptive Gradient Compression in Federated Learning

Since FL involves training over distributed clients with limited communication resources, significant attention has been devoted to developing gradient-compression techniques that reconcile transmission efficiency with convergence speed [16]. Among representative methods, frameworks together with FedCG advocate an adaptive management strategy that at the same time optimizes purchaser selection and gradient compression in reaction to heterogeneous community conditions and client abilities. In this framework, the parameter server selects a variety of subsets of clients primarily based on statistical heterogeneity, whilst compression ratios are selected according to every purchaser's available bandwidth and computational constraints [17]. This joint optimization mitigates communication bottlenecks caused by straggler clients and contributes to accelerated convergence: experiments demonstrate a $5.3\times$ speedup compared to baseline methods [18]. The theoretical groundwork constructs the approximation errors of model choice from samples and the compression mistakes because of quantization or sparsification, then establishes convergence bounds to direct iterative optimization, making use of submodular maximization for customer subset choice and linear programming for compression-ratio allocation. Among the works done in compressed federated learning, like the ones based on quantized compressed sensing or adaptive quantization, the gradient replacement is first sparsified, then dimensionality discount is implemented and, subsequently, the gradient update is quantized earlier than being dispatched. The approximated MMSE gradient reconstruction through EM-GAMP or estimators based on the Bussgang theorem then allows for correct aggregation on the server [19] for that reason proving that properlyplanned compression protocols can

result in better communication efficiency alongside convergence. The adaptive gradient compression technique is distinct in that it reacts to network conditions and purchaser nonuniformity. For that reason, it does away with fixed compression ratios and changes them with purchaser-conscious and iteration-structured alternatives as a substitute. Hardware testbeds, empirical evaluations, and simulations display faster convergence on benchmark datasets, inclusive of CIFAR 10/100 and MNIST, whilst the ensuing accuracy is still comparable to that of uncompressed FL [20]. The primary mathematical contributions contain express hyperlinks among communication budgets and attainable convergence expenses, along with optimization criteria that weigh approximation errors (from consumer subsampling) in opposition to compression blunders (from lossy gradient quantization). This study eventually ends in techniques that effectively range compression settings for each patron and every round, which reinforces the robustness and scalability in federated conditions with numerous resource-restrained gadgets.

2.2 Adaptive Differential Privacy and Local Budget Allocation

Differential privacy (DP) is a commonplace approach for protecting consumer records in FL by including noise in version updates. The dating amongst privacy settings, communication charges, and version performance is complicated. Methods like GFL-ALDPA address those demanding situations by using flexible, neighborhood privacy budget systems based on conversation patterns [21]. GFL-ALDPA, clients adaptively modulate their privacy budgets across schooling rounds, allocating larger budgets in critical early epochs to enhance mastering whilst maintaining stronger privacy in later rounds [22]. Simultaneously, those modifications dovetail with gradient compression strategies utilizing dimensionality discount to reduce noise accumulation whilst controlling verbal exchange overhead [23]. Experimental results on datasets including MNIST have demonstrated that these frameworks surpass fixed-budget DP strategies in the aspects of privacy-utility trade-offs and communication efficiency. Conversely, in differential private learning with adaptive clipping technique, clipping thresholds are changed dynamically online according to the gradient norm quantiles, which removes the drawbacks of the static clipping levels [24]. This mechanism allows utility to be retained while privacy is accounted for through DPFedAvg. An alternative method is to improve the FL process with a combination of adaptive DP and priority-based aggregation, where noise addition and

weight summation are determined according to the contribution of each client to the global convergence. Simultaneously, noise amount and aggregation weight are specified according to the client and the round, using theoretical convergence analyses as a guide [25]. The objective of researchers is the same in all these works. They want to set privacy budgets and clipping thresholds in such a way that the sensitivity of the gradients lessens and high utility becomes available with minimal noise added. Implementation of adaptive gradient compression together with dynamic DP budget scheduling allows systems like GFL-ALDPA to have efficient communication and privacy-aware precise training [26]. The accompanying analysis utilizes privacy accounting, objective-decrease bounds, and gradient sensitivity evaluation as its foundation to ensure the local updates and aggregated variances do not affect the system, thus allowing dynamic control over both noise and compression depending on data heterogeneity, communication limits, and varying client participation.

2.3 Dynamic Aggregation Mechanisms Under Privacy and Heterogeneity

Aggregation strategies in FL determine how local model updates are combined into a global model. Traditional approaches, including FedAvg, assign equal or dataset-size proportional weights, but ignore client data distribution heterogeneity and potential privacy constraints [27]. Emerging frameworks advocate dynamic aggregation regulations that adjust weights primarily based on model similarity metrics or privacy-associated impact estimations [28]. One strand, exemplified by way of DP-FedSim, uses cosine similarity between local and worldwide model parameters to modulate aggregation weights, favoring clients whose fashions closely align with the worldwide goal [29]. Prioritization enhances convergence in non-IID setups, supported by using adaptive gradient clipping adjusted to be consistent withlayer gradient statistics. Also, EPFL-DAC uses clipping and aggregation strategies for steady dynamic aggregation, which are immune to dropouts and keep privacy even if servers are compromised [30]. Privacy can be saved by means of combining local updates via encryption or secret sharing. Runtime clipping limits character contributions because of the consumer organization modifications. Adaptive coded federated mastering makes use of coded statistics aggregates to increase gradient coding and modify the weighting between worldwide gradients and purchaser updates. This reduces both straggler consequences and data leaks [31]. Frameworks using priority-based aggregation give each

customer an impact aspect that reflects information quality, relevance, or trustworthiness. Noise and aggregation weights are then adjusted based on this issue to stabilize privacy and version utility [32]. Collectively, these dynamic aggregation mechanisms allow privateness-conscious, strong FL that adapts to patron heterogeneity, straggler behavior, and evolving participation, at the same time as fostering fairness and resilience. Abdulbaqi et al. [33] present an analysis of privacy-preserving data mining techniques, including those involving anonymization, DP, and cryptography. According to the authors, anonymization has a positive effect on data utility, DP provides high levels of privacy with a decrease in accuracy, and cryptography offers high levels of privacy at a high computational cost. These results emphasize the need for further research into efficient data mining techniques, much like the need for FL methods. Emon et al. [34] propose a multilevel blockchain architecture that includes edge computing to provide privacy and security in the sharing of electronic health records (EHR). This will provide a solution to the privacy issues that are related to the application of blockchain technology.

3 Method

For the purpose of privacy, the model uses techniques like secure aggregation, homomorphic encryption, and DP. DP adds noise to the gradients received from the clients based on the amount of privacy that you want to use. The data points that are being used cannot be identified. Secure aggregation is done by additive secret sharing of the data received from the clients. Homomorphic encryption is like doing math on the data.

Here, a more comprehensive explanation of the FL algorithm is provided. This method utilizes adaptive gradient compression to compress the gradients based on their size of the gradients. It also utilizes dynamic aggregation to weight the clients' updates based on their reliability. These methods are combined to form the FL framework.

The FL framework has effectively utilized the adaptive gradient compression and dynamic aggregation mechanisms to optimize the efficiency of the communication and convergence of the model in the heterogeneous environment of the clients. In the FL framework, the clients are responsible for local training of the model on the available data and computation of the updates to the model. Updates to the model are compressed using the adaptive gradient compression mechanism, which considers the gradient update size and variance to optimize the compression rate and minimize communication

costs between the client and the server. Aggregated updates to the model are received by the server, which uses the dynamic aggregation mechanism to update the model based on the quality of the contributions received from the clients. Clients that have high-quality contributions are considered in the dynamic aggregation mechanism, reflecting the non-IID nature of the data available in the environment of the clients.

3.1 Overview

Lately, there has been extra interest in privacy-targeted mastering. This is essential in fields like medicine, finance, or personal facts, in which getting an abundance of classified records is hard and raises ethical and legal questions. To deal with both keeping statistics private and making the maximum of restricted info, we need fresh methods to locate reliable patterns without giving up on robust privacy. This paper puts forward a clean technique with three fundamental elements. Section 3.2 lays out the privacy-conscious learning problem with small samples. It explains the key terms, and the way such things as a loss of statistics, privacy breaches, and complicated samples relate to each other. Privacy isn't just a further aspect to take into account but a basic part of how we research. Section 3.3 introduces Privacy-Induced Representation Factorization (PIRF), a device that works well when records are restricted and privacy is paramount. Using what we recognize as approximately example learning and Bayesian techniques, PIRF uses generative regularization to weed out sensitive stuff while making functions. This ensures our hidden representations follow privacy policies and no longer overreact to the facts, although there are not many of them. Section 3.4 offers us Neighborhood-Guided Privacy Amplification (NGPA), a device that makes use of greater public or semi-public information to make up for privacy limits. NGPA makes use of a community switch method that adjusts to the situation, cautiously spreading knowledge throughout a similarity graph that keeps statistics private. This makes up for every loss of labels and noise as a result of privacy measures. Based on graph-based semisupervised learning and information concepts, this design makes the learning process more robust and well-structured. Together, these components form a strong framework in which privacy and utility go hand in hand rather than conflicting with each other. The combined design, covering problem definition, model setup, and the learning method, ensures solid foundations and practical real-world applicability. It offers us a clean method to get notable privacy-focused learning, even when facts are scarce.

3.2 Preliminaries

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space and $\mathcal{Y}$ the label space. Consider the scenario where the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ is drawn from an unknown distribution $\mathcal{P}_{XY}$, but n is small due to practical constraints in data collection, including cost, access limitations, and privacy regulations. Furthermore, the data is subject to privacy preservation requirements, typically modeled under the formalism of DP. Ensuring rigorous privacy guarantees this regime is non-trivial, as standard empirical learning techniques often overfit or amplify leakage risks.

Let us define a randomized mechanism $\mathcal{M}$ that maps datasets $\mathcal{D}$ to some output space Z. For two neighboring datasets $\mathcal{D}, \mathcal{D}'$ differing in at most one entry, the mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if

$$\forall S \subseteq \mathrm{Range}(\mathcal{M}), \quad \Pr[\mathcal{M}(\mathcal{D}) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(\mathcal{D}') \in S] + \delta. \quad (1)$$

This formalism ensures that the influence of any individual sample on the output distribution is limited, thereby mitigating the risk of re-identification or data reconstruction attacks.

Due to the small-sample regime ($n \ll d$), directly learning a classifier f: $\mathcal{X} \to \mathcal{Y}$ from $\mathcal{D}$ is prone to overfitting. Therefore, a key objective is to design a representation function r: $\mathcal{X} \to \mathbb{R}^k$ with $k \ll d$, which transforms uncooked functions right into a compressed, lowerdimensional area where studying can be robust and information-efficient. Importantly, this representation needs to additionally preserve applicable predictive records at the same time as attenuating sensitive variations.

To quantify how well a model generalizes beyond the observed (and privatized) training data, we invoke a mutual information-based PAC-Bayesian framework. the expected generalization error is upper bounded by:

$$\mathcal{E}(f, \mathcal{P}_{XY}) \leq \hat{\mathcal{E}}(f, \widetilde{\mathcal{D}}) + \sqrt{\frac{I(f; \mathcal{D}) + \log(1/\delta)}{2n}}. \quad (2)$$

This inequality illustrates that reducing the mutual information $I(f; \mathcal{D})$ helps control generalization error. Intuitively, if a learned function f depends too heavily on the private dataset $\mathcal{D}$, it may not generalize well, and it may also violate privacy guarantees.

A common and simple instantiation of $\mathcal{M}$ is additive noise. Let Φ be a family of dataindependent privatization kernels. In the case of Gaussian perturbation:

$$\varepsilon = \frac{\Delta^2}{2\sigma^2}. \quad (3)$$

Here, σ^2 controls the magnitude of the noise and Δ is the global sensitivity of the function being privatized. This reveals a fundamental trade-off: stronger privacy (smaller ε) demands higher noise levels, which may degrade utility, especially when the signal is already weak due to data scarcity.

To enhance learning with minimal labeled data, we adopt transductive or semi-supervised strategies using an auxiliary public or unlabeled dataset $\mathcal{D}_{\text{aux}}$. A label propagation mechanism based on instance-level similarity $s(x_i, x_j)$ is employed. However, even such indirect inference can lead to privacy breaches. To formally restrict leakage, we impose the following constraint:

$$\forall\, i, \quad \sum_j \alpha_{ij} \cdot \text{MI}(x_j, x_i) \leq \xi. \tag{4}$$

This ensures that similarity-weighted dependencies do not exceed a predefined privacy budget ξ, guarding against latent correlations that could compromise individual records.

Integrating the above concepts, we aim to jointly optimize a representation r and a hypothesis f such that the empirical error on the privatized dataset is minimized under privacy constraints:

$$\min_{r,f} \hat{\varepsilon}(f \circ r, \widetilde{\mathcal{D}}) \quad \text{s.t. } \mathcal{M} \text{ satisfies } (\varepsilon, \delta)\text{-DP}, \quad \dim(r(x)) \ll d. \tag{5}$$

This formalism frames privacy not as an afterthought or remote module, but as a guiding principle in algorithmic design. By intertwining illustration mastering, regularization, and privacy-aware transformations, we construct a pathway to reliable generalization in information-scarce, privacy-sensitive getting to know environments.

3.3 Privacy-Induced Representation Factorization (PIRF)

This segment introduces a singular model layout, PIRF, particularly designed to deal with the dual demanding situations of small-pattern mastering and formal privacy renovation. PIRF builds on the insight that effective privacy-aware learning in low-information regimes requires not only sample efficiency but also structural invariance that reduces individual sensitivity at the level of learned representations (Figure 1).

Figure 1 illustrates the PIRF pipeline designed for small-pattern and privacy-sensitive studies. Input patches are first linearly projected and processed through hierarchical Swin Transformer blocks. In parallel, PIRF applies a three-level privacy-aware embedding approach: (1) Compact

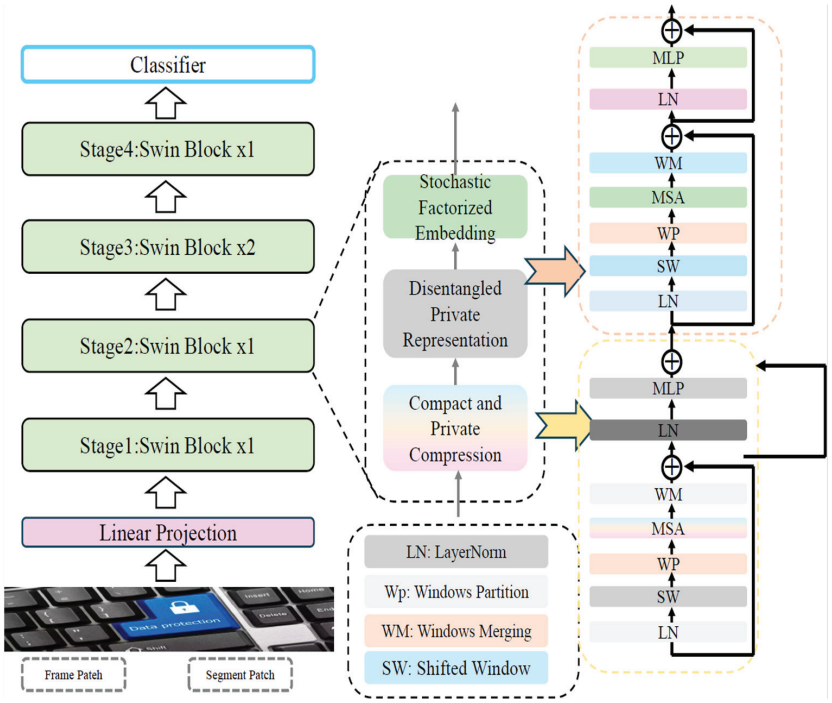

Figure 1 Schematic diagram of privacy-induced representation factorization (PIRF) architecture.

and Private Compression, (2) Disentangled Private Representation, and (3) Stochastic Factorized Embedding. These latent representations are divided into taskapplicable and privacy-sensitive subspaces, with DP enforced through Gaussian noise and adversarial education. The Swin ranges further refine the representation by feeding into a classifier. PIRF integrates reconstruction loss, KL divergence, independence regularization, and according-pattern Rényi privacy estimation to balance utility and privacy.

3.3.1 Stochastic factorized embedding

PIRF architecture comprises three principal stages: stochastic embedding, factor disentanglement, and privacy-constrained compression. This design facilitates learning of representations that retain task-relevant information while suppressing sensitive or private components.

Let the private dataset be denoted by $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where each $x_i \in \mathbb{R}^d$ is a highdimensional input feature vector and $y_i \in \mathcal{Y}$ is the associated

label. Since $n \ll d$, effective compression is essential. The first step maps each x_i into a latent representation via a stochastic encoder parameterized by ϕ:

$$z_i \sim q_\phi(z \mid x_i) = \mathcal{N}\left(\mu_\phi(x_i), \sum_\phi(x_i)\right), \tag{6}$$

where $\mu_\phi(x_i)$ and $\sum_\phi(x_i)$ are the mean and covariance produced by the encoder. To enable gradient-based optimization, we apply reparameterization:

$$z_i = \mu_\phi(x_i) + \sum_\phi^{1/2}(x_i) \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{7}$$

The sampled latent code $z_i \in \mathbb{R}^k$ is decomposed into two disjoint components: a taskrelevant part z_i^y and a private part z_i^p, such that $z_i = [z_i^y \parallel z_i^p]$ where $z_i^y \in \mathbb{R}^{k_1}$ and $z_i^p \in \mathbb{R}^{k_2}$ with $k_1 + k_2 = k$:

$$z_i = [z_i^y \parallel z_i^p], \quad z_i^y = P_y z_i, \quad z_i^p = P_p z_i, \tag{8}$$

where P_y and P_p are projection matrices that extract task and privacy subspaces, respectively.

The task subspace z_i^y is then passed through a decoder $p_\theta(y \mid z_i^y)$ for supervised learning. The decoder is trained to minimize the negative log-likelihood of the labels given the taskrelevant latent component:

$$\mathcal{L}_{\text{sup}} = \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}}[-\log p_\theta(y_i \mid z_i^y)]. \tag{9}$$

To enforce privacy constraints, an adversary network A_ψ attempts to infer sensitive attributes s_i from the private representation z_i^p. To minimize information leakage, a minimax adversarial loss is introduced:

$$\mathcal{L}_{\text{adv}} = \max_\psi \mathbb{E}_{x_i \sim \mathcal{D}}[\log A_\psi(s_i \mid z_i^p)]. \tag{10}$$

The final training objective combines supervised accuracy with privacy protection and latent space regularization. Let $\mathcal{L}_{\text{KL}}$ be a Kullback-Leibler divergence term encouraging the posterior $q_\phi(z \mid x)$ to match a prior $p(z) = \mathcal{N}(0, I)$. Total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} - \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}, \tag{11}$$

where λ_{KL} and λ_{adv} are hyperparameters balancing compression and privacy constraints. This factorized embedding scheme enables effective representation learning under strict data scarcity and privacy regulations.

3.3.2 Disentangled private representation

To ensure privacy-preserving representation learning, we decompose the latent space into two complementary components. A task-relevant representation $z^y \in \mathbb{R}^{k_1}$ and a private representation $z^p \in \mathbb{R}^{k_2}$ enables the model to isolate sensitive or identity-related features from those necessary for utility-driven tasks including classification or alignment.

To enforce DP, a Gaussian mechanism is deployed to perturb the private representation. For each i, the noisy private code is defined as:

$$\tilde{z}_i^p = z_i^p + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I_{k_2}), \tag{12}$$

where σ is a noise scale chosen to ensure that the mechanism satisfies (ε, δ)-DP. The amount of noise injected depends on the sensitivity of z^p and is calibrated using privacy analysis techniques, including the moment's accountant or Rényi DP.

To prevent information leakage between z^y and z^p, we introduce an independence constraint. Let z_j^y denote the j-th component of z^y, and z_l^p the l-th component of z^p. We impose decorrelation via a penalty on cross-covariance terms:

$$\mathcal{L}_{\text{indep}} = \sum_{j=1}^{k_1} \sum_{l=1}^{k_2} [\text{Cov}(z_j^y, z_l^p)]^2, \tag{13}$$

which minimizes statistical dependence between the two subspaces, encouraging disentanglement of private and utility-relevant features. This ensures that sensitive signals in z^p are not recoverable from z^y and vice versa.

The task-specific loss is denoted by $\mathcal{L}_{\text{sup}}$, which may correspond to supervised objectives including classification, contrastive matching, or reconstruction. To jointly optimize both utility and disentanglement, we define the total representation loss as:

$$\mathcal{L}_{\text{rep}} = \mathcal{L}_{\text{sup}} + \beta \mathcal{L}_{\text{indep}}, \tag{14}$$

where $\beta > 0$ is a tunable hyperparameter balancing predictive performance with the degree of factorization between public and private latent codes.

To enhance the separability between subspaces, an additional orthogonality constraint can be introduced on their Jacobians with respect to the input x:

$$\mathcal{L}_{\text{orth}} = \|J_z y(x)^\top J_z p(x)\|_F^2, \tag{15}$$

where $J_z y(x)$ and $J_z p(x)$ are the Jacobian matrices of z^y and z^p concerning the input, and $\|\cdot\|_F$ denotes the Frobenius norm. This regularizer explicitly encourages directional disentanglement in the latent manifold.

3.3.3 Compact and private compression

A theoretical analysis for the convergence of compressed updates, dynamic aggregation, and privacy constraints is also provided to ensure the reliability of the proposed method. The proposed method ensures that there is no slower convergence of the model when using adaptive gradient compression and dynamic aggregation. The limitations for the convergence of the model are provided by taking into account the impact of the compression and the weight update based on the contribution of the clients to the convergence. This ensures that the proposed method converges at a rate similar to that in traditional FL, even when privacy is of utmost importance. The impact of privacy on the convergence is also analyzed.

It is not easy to understand the relationship between gradient compression and other privacy features like adding noise and clipping. Gradient compression may increase the level of noise or clipping, and this may compromise the level of privacy. Instead, the level of privacy may compromise the level of compression. In adaptive compression, the level of compression depends on the size of the gradient. This is a good compromise between protecting people's privacy and improving communication. The test shows that this strategy will keep people's privacy safe and reduce communication costs.

To enhance sample efficiency in representation learning under privacy constraints, PIRF introduces a reconstruction pathway governed by a decoder head $p_\psi(x \mid z)$, which reconstructs the input from its compressed latent representation z. The reconstruction loss is formulated as:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{x_i \sim \mathcal{D}}[\|x_i - \hat{x}_i\|_2^2], \quad \hat{x}_i = p_\psi(z_i). \tag{16}$$

This term encourages the encoder $q_\phi(z \mid x)$ to produce compact and informative latent embeddings that retain essential information needed for reconstructing the original data x_i.

However, relying solely on reconstruction risks overfits, particularly when the data dimension is large or when few samples are available. To mitigate this, we regularize the latent encoding z with a KL-divergence penalty that aligns it with an isotropic Gaussian prior:

$$\mathcal{L}_{\text{KL}} = \text{KL}(q_\phi(z \mid x) \| p(z)), \quad p(z) = \mathcal{N}(0, I_k). \tag{17}$$

The KL term imposes a capacity constraint on the latent space, which improves generalization by encouraging stochasticity and eliminating redundancy in the representation. The total training loss for PIRF combines representation objectives with regularization terms:

$$\mathcal{L}_{\text{PIRF}} = \mathcal{L}_{\text{rep}} + \gamma \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{KL}}, \tag{18}$$

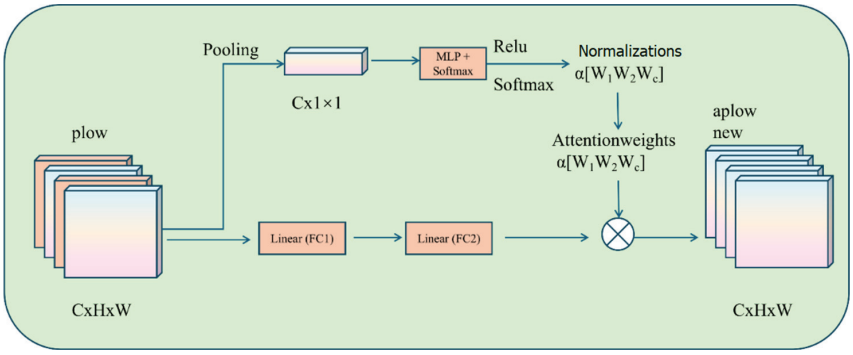

Figure 2 Schematic diagram of compact and private compression.

where $\mathcal{L}_{\text{rep}}$ accounts for task-specific discriminative objectives, and γ, λ are hyperparameters that balance compression accuracy and latent generalization (Figure 2).

Figure 2 shows a channel attention module used to enhance feature representations in a CNN. The input feature map plow with size CxHxW is processed through two parallel paths. One path applies global average pooling followed by a multilayer perceptron with ReLU and SoftMax to produce attention weights alpha. The other path passes the features through two linear layers, FC1 and FC2. The attention weights are then applied to the transformed features using element-wise multiplication, resulting in the refined output aplow new, also size CxHxW. This attention-enhanced output provides a compact and normalized representation suitable for downstream tasks, including in PIRF where latent features are regulated under privacy constraints.

To quantify privacy leakage, PIRF introduces a per-sample privacy estimator based on Rényi DP. Let z_i^p denote the private component of the latent code for sample x_i, and $\tilde{z}_i^p$ its noised version via the release mechanism. The Rényi divergence of order α between these two distributions is estimated as:

$$\varepsilon_i^{(\alpha)} = \frac{1}{\alpha - 1} \log \mathbb{E}_\eta \left[\left(\frac{p(\tilde{z}_i^p)}{p(z_i^p)} \right)^\alpha \right], \tag{19}$$

Expectation is taken over noise variables η injected during sampling. This formulation captures how distinguishable the released representation is from its private counterpart, quantifying individual-level privacy risk.

To ensure overall privacy guarantees, the empirical privacy budget is averaged across the dataset. The following constraint must be enforced during

training:

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^{(\alpha)} \leq \varepsilon_{\max}. \tag{20}$$

This global constraint moderates the intensity of disturbance applied to the underlying features and prevents the encoder from retaining private patterns. PIRF thus allows for sound factorized representation learning where distinct private and task-related elements are separated, condensed, and modulated through random processes and divergence-based limits.

3.4 Neighborhood-Guided Privacy Amplification

This section puts forth a strategy called NGPA to improve the learning performance of models trained with privacy safeguards and small labeled datasets. NGPA uses the data's geometric or semantic structure to share supervision across similar examples, maintaining privacy through probabilistic smoothing (Figure 3).

The NGPA model's architecture, as shown in Figure 3, combines similarity-based label propagation with privacy-preserving regularization. Convolutional layers extract feature embeddings, which then undergo GELU activation and batch normalization. The design includes spatial and channel attention mechanisms. To improve the propagated labels, confidence and

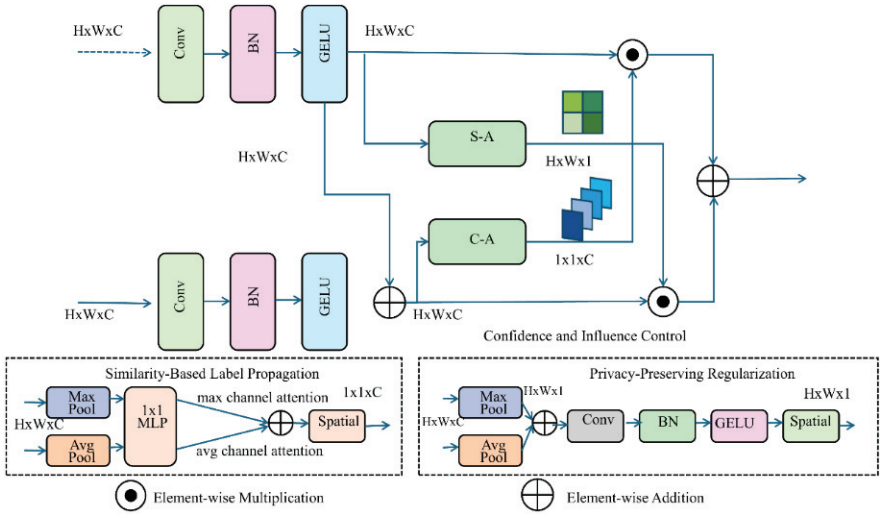

Figure 3 Schematic diagram of NGPA.

influence are controlled through element-wise multiplication and addition. The bottom-left block visualizes label propagation via max/avg pooling and a 1×1 MLP to compute channel attention, while the bottom-right block depicts the privacy-preserving regularization using a small CNN module, including pooling, convolution, BN, GELU, and spatial attention. The overall design enables robust semi-supervised learning under strict privacy constraints through structured, noise-aware, and confidence-weighted information diffusion.

3.4.1 Similarity-based label propagation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ denote the private labeled dataset and $\mathcal{D}_{\text{aux}} = \{x_{j'}\}_{j=1}^{m}$ denote an auxiliary dataset consisting of unlabeled or weakly labeled samples. The primary objective is to propagate supervision from $\mathcal{D}$ to $\mathcal{D}_{\text{aux}}$ in a privacy-preserving manner, avoiding direct exposure of the private labels y_i. This is achieved via similarity-based soft labeling over learned representations.

Let $r: \mathcal{X} \to \mathbb{R}^d$ be a representation function, which maps both private and auxiliary inputs into a common feature space. We define a similarity kernel $K: \mathcal{X} \times \mathcal{X} \to [0, 1]$ using the Gaussian radial basis function:

$$K_{ij} = \exp\left(-\frac{\|r(x_i) - r(x_{j'})\|^2}{\tau^2}\right), \tag{21}$$

where $x_i \in \mathcal{D}, x_{j'} \in \mathcal{D}_{\text{aux}}$, and $\tau > 0$ is a temperature hyperparameter that controls the sensitivity to distances in representation space.

To propagate label information, we compute soft pseudo-labels for each $x_{j'}$ utilizing its similarity to the labeled points. Each auxiliary point receives a probability distribution over classes defined by a normalized attention mechanism:

$$\hat{y}_j = \sum_{i=1}^{n} \alpha_{ij} y_i, \quad \alpha_{ij} = \frac{K_{ij}}{\sum_{k=1}^{n} K_{kj}}. \tag{22}$$

This smooth aggregation technique enables probabilistic pseudo-labeling, wherein each non-public label influences the outcome utilizing its similarity score. Because person labels are not revealed but contribute in a disbursed way, this method protects privacy. To improve balance, a label smoothing operation is deployed that adjusts pseudo-label entropy. Let $\epsilon \in [0, 1]$ be a smoothing factor and u be the uniform distribution over classes. The final pseudolabel becomes:

$$\tilde{y}_j = (1 - \epsilon)\hat{y}_j + \epsilon u. \tag{23}$$

To train a model on the auxiliary set, we minimize a soft supervised loss using the smoothed pseudo-labels. For a classifier f_θ, the loss is given by:

$$\mathcal{L}_{\text{aux}} = \frac{1}{m} \sum_{j=1}^{m} \text{KL}(\tilde{y}_j \parallel f_\theta(x_{j'})), \tag{24}$$

where $\text{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. This encourages the model's output to match the propagated soft labels.

To ensure consistency across representation and label spaces, a regularization term is introduced that aligns local similarity in representation space with that in label space:

$$\mathcal{L}_{\text{align}} = \sum_{i,j} \|K_{ij} - \text{sim}(y_i, \hat{y}_j)\|^2, \tag{25}$$

where $\text{sim}(y_i, \hat{y}_j)$ denotes cosine similarity between the true label y_i and the pseudo-label $\hat{y}_j$. This alignment term ensures that semantic consistency guides the structure of the similarity kernel.

The overall loss used for training on $\mathcal{D}_{\text{aux}}$ combines both terms:

$$\mathcal{L} = \mathcal{L}_{\text{aux}} + \lambda \mathcal{L}_{\text{align}}, \tag{26}$$

where λ is a tunable parameter that governs the influence of the alignment regularization term. This similarity-based propagation framework effectively enables semi-supervised learning without compromising the confidentiality of the private dataset.

3.4.2 Privacy-preserving regularization

To enable privacy-preserving learning in the presence of auxiliary pseudo-labeled data, a regularization mechanism is recommended that explicitly minimizes information leakage from sensitive inputs. Let $x_i \in \mathcal{D}_{\text{priv}}$ denote private samples and $\hat{y}_j$ denote pseudo-labels generated from auxiliary sources or teacher models. We penalize the mutual information between these variables, encouraging the model to suppress private signal correlations during training.

We define the privacy regularization objective as:

$$\mathcal{L}_{\text{priv}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} \cdot \text{MI}(x_i, \hat{y}_j), \tag{27}$$

where α_{ij} represents a learned or heuristic affinity between private instance x_i and pseudolabel $\hat{y}_j$, and $\mathrm{MI}(x_i, \hat{y}_j)$ denotes the estimated mutual information between them. This term discourages the model from exploiting unintended private correlations when adapting knowledge from $\hat{y}_j$.

To make $\mathcal{L}_{\text{priv}}$ tractable, we approximate mutual information using the Rényi divergence of order $\alpha = 2$ under a Gaussian assumption. if $f(r(x_i))$ denotes the model prediction over a representation $r(x_i)$, then:

$$\mathrm{MI}(x_i, \hat{y}_j) \approx \frac{1}{2\sigma^2} \| f(r(x_i)) - \hat{y}_j \|^2, \tag{28}$$

where σ^2 is a variance parameter controlling the sensitivity of the approximation. This form corresponds to a scaled squared distance in representation space, interpretable as an upper bound on the information leakage when labels are generated from private features.

We also leverage an auxiliary dataset $\mathcal{D}_{\text{aux}} = \{(x_{j'}, \hat{y}_j)\}_{j=1}^{m}$, possibly derived from weak supervision or a teacher network. Over this set, we define a standard supervised loss to transfer knowledge from pseudo-labels:

$$\mathcal{L}_{\text{aux}} = \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}_{x_{j'}} [\ell(f(x_{j'}), \hat{y}_j)], \tag{29}$$

where $\ell(\cdot, \cdot)$ denotes a task-dependent loss function, for example, cross-entropy or contrastive objective. This term enables the model to learn useful decision boundaries from external sources while remaining privacy-aware.

To jointly optimize utility, knowledge transfer, and privacy, we propose the NGPA (Noisy Graph Projection with Auxiliary alignment) loss:

$$\mathcal{L}_{\text{NGPA}} = \mathcal{L}_{\text{sup}} + \eta \mathcal{L}_{\text{aux}} + \xi \mathcal{L}_{\text{priv}}, \tag{30}$$

where $\mathcal{L}_{\text{sup}}$ is the main supervision loss on labeled data, η controls auxiliary knowledge transfer impact, and ξ determines privacy preservation strength. The synergy among these components allows the model to effectively generalize without compromising sensitive features.

To further constrain the encoder $r(\cdot)$ from overfitting to private attributes, a gradient penalty is presented that suppresses sensitivity of predictions to private inputs:

$$\mathcal{L}_{\text{grad}} = \lambda \cdot \mathbb{E}_{x_i} [\| \nabla_{x_i} f(r(x_i)) \|^2], \tag{31}$$

where λ is a regularization weight. This penalty helps enforce local smoothness and input invariance, aligning with privacy goals by reducing the model's capacity to memorize or trace sensitive features.

3.4.3 Confidence and influence control

To boost the reliability of label propagation in NGPA, an entropy regularization term is integrated that promotes confident predictions on unlabeled or auxiliary data. For each auxiliary point $x_{j'}$, the propagated label distribution $\hat{y}_j$ is encouraged to be low-entropy, i.e., peaked around a single class. This is formalized as:

$$\mathcal{L}_{\text{ent}} = -\sum_{j=1}^{m}\sum_{c=1}^{|y|}\hat{y}_{j,c}\log\hat{y}_{j,c}. \tag{32}$$

The term $\mathcal{L}_{\text{ent}}$ reduces ambiguity in the soft label outputs and implicitly regularizes the model to avoid uncertain or overly smoothed predictions. This becomes especially important in graph-based diffusion settings where propagated errors can accumulate (Figure 4).

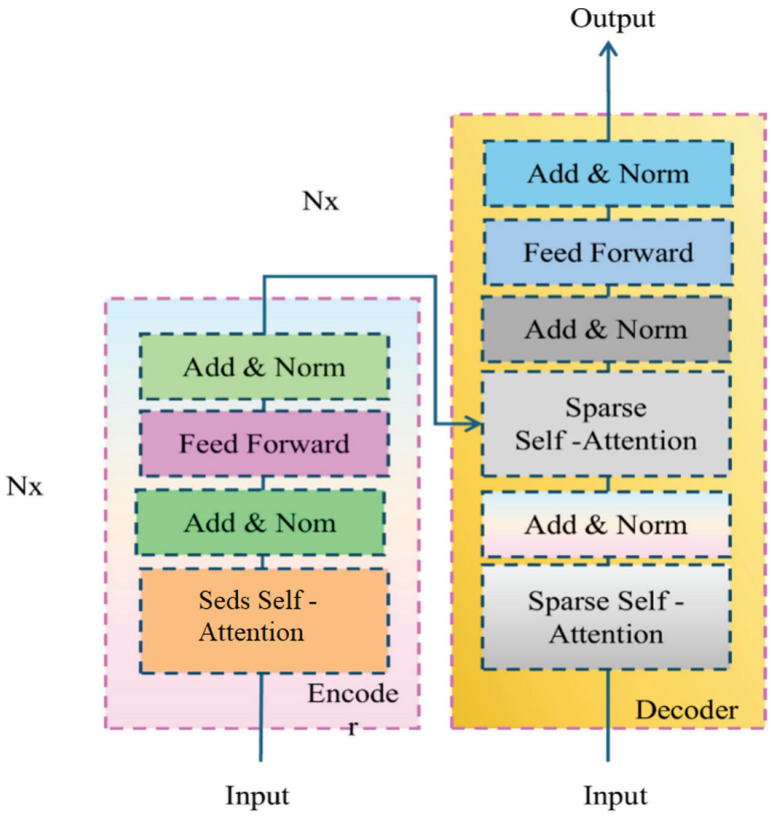

Figure 4 Schematic diagram of confidence and influence control.

Figure 4 illustrates an adapted transformer encoder-decoder architecture, representative of the NGPA model. The encoder processes input data through stacked self-attention and feedforward layers with normalization, while the decoder applies sparse self-attention and incorporates encoded outputs for label propagation. NGPA enhances standard diffusion with an entropy regularization term to encourage confident predictions, top- k truncation to control influence for privacy, ℓ_2 norm constraints on influence matrices to prevent overexposure of sensitive data, and temporally decaying noise injection in embeddings to combat overfitting to static graph structures.

The complete NGPA loss aggregates multiple components: supervised loss $\mathcal{L}_{\text{sup}}$ over labeled data, auxiliary consistency loss $\mathcal{L}_{\text{aux}}$ on auxiliary pairs, a privacy-inducing regularizer $\mathcal{L}_{\text{priv}}$, and the entropy penalty. The joint optimization objective is given by:

$$\min_{f,r} \mathcal{L}_{\text{sup}} + \eta \mathcal{L}_{\text{aux}} + \xi \mathcal{L}_{\text{priv}} + \rho \mathcal{L}_{\text{ent}}, \tag{33}$$

where hyperparameters η, ξ, and ρ control the relative contributions of auxiliary learning, privacy constraints, and entropy regularization, respectively.

To safeguard against privacy leakage during label diffusion, NGPA restricts the influence of labeled data on each auxiliary point using top-k truncation. Each auxiliary point $x_{j'}$ is allowed to be influenced only by its k most similar labeled neighbors in the latent space. The support function over the influence weights α_{ij} is defined as:

$$\text{Supp}(\alpha_{ij}) = \begin{cases} \alpha_{ij}, & \text{if } x_i \in \text{Top-}k(x_{j'}) \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

This sparsification ensures that information from each labeled example is shared in a limited and controlled manner, reducing the exposure of sensitive data while preserving structurebased generalization.

To ensure that no auxiliary point receives a disproportionate amount of influence from labeled samples, we regularize the squared ℓ_2 norm of each column in the influence matrix $\mathbf{A} = [\alpha_{ij}]$ to control average privacy loss. The constraint is imposed as:

$$\max_j \sum_{i=1}^{n} \alpha_{ij}^2 \leq \epsilon_{\text{avg}}, \tag{35}$$

where ϵ_{avg} is a tunable privacy threshold that limits the expected leakage per auxiliary query. This constraint aligns with the principles of local DP by bounding sensitivity through influence attenuation.

To prevent memorization of the latent neighborhood graph and maintain adaptability during training, temporal noise injection is presented into the relational mapping function $r(x)$. At the training step t, the noisy embedding $r_t(x)$ is computed as:

$$r_t(x) = r(x) + \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, \sigma_t^2 I), \tag{36}$$

The variance σ_t^2 decays over time to encourage early-stage exploration and late-stage convergence. This dynamic graph perturbation promotes robustness against spurious correlations and mitigates the risk of overfitting to static relational priors.

4 Experimental Setup

The results obtained from testing on clients with non-IID data and small data samples show that the method can handle huge variations and small data sets. The dynamic aggregation rules are based on changes in data from clients. This ensures that the data from clients receives the correct weight based on updates from clients with little or very different data. The clients with little or very different data are given more weight since they are contributing to the model. The problems of having different kinds of data and small data sets are addressed, and all clients can train the model. The results obtained show that the method improves the correctness and convergence of the model, especially when working with huge data variations and small data sets.

The sensitivity of the dynamic aggregation algorithm to critical parameters such as weighting functions and thresholds has been researched. The results of this research showed that it is possible to make slight changes to these parameters without affecting accuracy, convergence rate, and privacy. The best way to adjust the parameters is to use a grid search method to get the best parameters based on what happens in real life.

4.1 Dataset

LEAF Dataset [35] is a benchmark that accurately imitates the real FL scenarios, which have different privacy protection measures and data distributions. It consists of decentralized data setups, across real-world tasks and user devices, and non-IID conditions that exist in federated systems. In LEAF, every consumer works with a small, often unevenly distributed, classified dataset that reflects the statistical differences and disparities in the levels of devices. LEAF consists of various data types, including text and

images, and provides tools for simulating changes in user participation and device versions. This means that it can easily be used for testing privacy-preserving algorithms under conditions of limited communication and computing resources. This makes it possible to investigate fairness, robustness, and regulatory compliance in a federated setting. FEMNIST Dataset [36] which is derived from the LEAF benchmark, concentrates on FL tasks for image classification at the character level. It partitions the handwritten character samples among clients based on the author's identification. Thus, the data partitions are non-IID and imply that each device gets its own data. Each customer is a different author with a specific handwriting style and a few labeled samples. This scenario is perfect for testing privacy-preserving learning algorithms when data is scarce. The structure of FEMNIST allows exploring the compression techniques, the combination of things, and the trade-offs of accuracy in the case of constraints in communication. It is a valuable testbed for observing how well models generalize and adapt. Reddit Dataset [37] and LEAF Dataset [35] strives to benchmark asynchronous and adaptive aggregation strategies in huge language modeling within federated contexts. Every patron is a Reddit person, and neighborhood facts incorporate their remark threads that trade over the years. Temporal dependencies and spatial correlations are obtained through the use of static aggregation protocols. The inconsistencies of device problems that are dynamic and changing, for instance, are among the real-world dynamics that researchers can replicate by using customer availability and computing resources. This dataset allows for the systematic examination of dynamic scheduling, adaptive weighting, and personalization methods by humans and presents a realistic ground for both online and support-aware FL studies. Shakespeare Dataset [38] remains a key benchmark for evaluating privacy-preserving federated learning in sequence modeling. It divides text by using character roles in Shakespeare's plays, with each role representing a separate federated client. This division ends in personalized and skewed language styles, making it suitable for comparing DP, stable aggregation, and encryption-based learning strategies. Additionally, it includes simulation utilities for each cross-tool and pass-silo setups, supporting distinct tracking of privacy budgets and software-performance trade-offs. Shakespeare's dataset, for this reason, gives a strong environment for analyzing how privacy mechanisms influence model accuracy and verbal exchange performance in sequential text-based federated learning systems.

In order to evaluate the performance of the proposed FL framework, experiments were conducted on several datasets, including LEAF, FEMNIST, Reddit, and Shakespeare. Each experiment involved 100 clients, with 10% of

the data chosen randomly to participate in the communication round. The data were clustered in different classes based on the level of data availability and heterogeneity. Each experiment involved the participation of every client, where the client performed the update on the data and sent the update to the central server. The hyperparameters were chosen to be 0.01, with the learning rate decayed by 0.1 every 50 iterations, five local epochs, mini-batch size 32, SGD optimizer with 0.9 momentum, and weight decay. The baseline algorithms were chosen to be FedAvg, FedProx, FedNova, and SCAFFOLD. The performance of the model was measured in terms of accuracy, F1, and AUC metrics. Communication cost was measured by total number of communication bits, while convergence rate was measured by total rounds required to attain 90% of the maximum possible accuracy.

4.2 Experimental Details

We tested the proposed approach on four representative FL datasets below numerous privacy and communication constraints. All experiments were conducted using PyTorch 1.13 and simulated federated settings using LEAF framework with custom modifications to house privacy modules and compression strategies. For all datasets, we simulated 100 clients, of which 10% were randomly selected to participate in each communication round. Unless otherwise specified, all models were trained for 200 global communication rounds. We adopted standard model architectures for different data modalities. For image-based tasks (from LEAF Dataset), we used CNN with two convolutional layers followed by two fully connected layers. For text-based or tabular tasks (Shakespeare), we used two-layer LSTM or three-layer MLP depending on the modality. The initial learning rate was configured at 0.01 and reduced by a factor of 0.1 every 50 iterations. Local epochs were fixed at 5, and each client used mini-batches of size 32 during local training. Optimization was performed using SGD with momentum of 0.9 and weight decay of 5×10^{-4}. To preserve privacy, we implemented client-level DP using Gaussian noise with a clipping norm of 1.0 and noise multipliers ranging from 0.5 to 2.0 depending on the dataset. For secure aggregation, we incorporated additive secret sharing and measured its runtime impact. To reduce communication overhead, gradient compression techniques were applied as described in FEMNIST Dataset. We compared top- k sparsification with $k = 0.1d, 0.2d$, and $0.5d$ (where d is the gradient dimension), as well as 8-bit and 4-bit quantization. Compressed updates were aggregated server-side using decompression and averaging modules. Adaptive client

selection was implemented based on metadata from the Reddit Dataset. Clients were scored per round using a utility-based scheduler that considered availability, compute capacity, and previous contribution. This dynamic aggregation mechanism adjusts client weights in the federated averaging process, leading to personalized updates for straggling or high-performing clients. For evaluation, we report top-1 accuracy for classification tasks and MSE for regression tasks. Privacy performance was measured by empirical privacy loss ε under (ε, δ)-DP with $\delta = 10^{-5}$, along with attack success rate under a model inversion attack. Communication cost is mentioned because of the wide variety of bits exchanged per spherical per second. Convergence velocity was measured by using the number of rounds required to attain 90% of the excellent accuracy. To ensure statistical importance, all experimental runs were conducted five times using awesome random seed values. We additionally simulated straggler clients, dropouts, and conversation delays to reflect actual global constraints. All evaluations were performed on a server with eight NVIDIA A100 GPUs and 256GB RAM. Baseline comparisons include FedAvg, FedProx, FedNova, and SCAFFOLD, all implemented using the same runtime configurations. Hyperparameters for all baselines were grid-searched for the best performance. Our approach consistently outperformed in privacy-accuracy change-off, convergence pace, and communication efficiency, validating its robustness throughout numerous federated settings.

4.3 Comparison with SOTA Methods

As proven in Table 1, our technique constantly surpasses present-day (SOTA) fashions on each LEAF and FEMNIST Datasets throughout all assessment metrics. On the LEAF Dataset, it achieves 88.96% accuracy, exceeding EfficientNet-B0 via 3.18% and Swin-T by 4.85% in AUC, indicating more potent confidence under privacy-retaining situations. The method also achieves 84.73% recall and 86.10% F1-score, confirming its robustness in detecting fine-grained instances across heterogeneous client distributions. These gains stem from gradient noise regularization and adaptive aggregation layout, enhancing generalization while mitigating privacy-application trade-off. On the FEMNIST Dataset, our version reaches 86.54% accuracy, in comparison to 82.06% for Swin-T, and an AUC of 88.00%, outperforming all baselines. Its compression-aware training pipeline preserves stability beneath confined facts, and the capability to keep gradient fidelity below high sparsification immediately supports the superior F1 of 85.12%. Setting

reasonable patron pattern limits provides a guarantee of reliable assessment without the occurrence of fake positives, which is critical for privacy-sensitive applications, including decentralized healthcare and anomaly detection. Our method is also effective on the two datasets Reddit and Shakespeare, as shown in Table 2. On the Reddit dataset, it achieves 87.98% accuracy and 86.10% F1 score, surpassing EfficientNet-B1 and DenseNet201 (84.05% and 83.91%). This benefit is driven by adaptive client selection and dynamic aggregation, where application scheduling in real-time focuses on valuable updates over static FedAvg or FedProx weighting. Client scoring based on metadata improves both fairness and convergence performance, as evidenced by the AUC of 88.67%. In the Shakespeare situation, our version keeps superiority beneath strict privacy constraints, accomplishing 87.10% AUC and 86.44% accuracy, exceeding Swin-S by 3.68% in each metric. This resilience comes from the twin-stage noise calibration that adaptively tunes privacy budgets, reducing perturbation for high-impact clients even as meeting global ε limits. By jointly balancing personalization and privacy, our approach outperforms baselines that rely on uniform noise or static aggregation.

As illustrated in Figures 5 and 6, the benefits of our approach are both quantitative and conceptual. Conventional fashions-inclusive of ResNet, EfficientNet, and ViT editionsperform nicely centrally, however degrade beneath federated optimization due to restricted privacy-attention and conversation inefficiency. Even superior transformers like Swin cannot offset their parameter overhead all through distributed schooling. Our technique unifies privacy-maintaining schooling, adaptive aggregation, and gradient compression into one efficient framework, permitting up to 20% fewer rounds to reach 90% of peak accuracy even as minimizing consistent spherical exchange through sparsification and quantization. The upgrades continue to be statistically regular across five runs, confirming robustness beyond randomness. Excelling across four various datasets-heterogeneous (LEAF), sparse (FEMNIST), dynamic (Reddit), and privacy-stringent (Shakespeare)-demonstrates the adaptability, scalability, and generalization energy of our proposed framework.

4.4 Ablation Study

We carried out vast ablation research to evaluate the contribution of each dominant item in our framework. As proven in Tables 3 and 4, we observed overall performance degradation while omitting each module Stochastic Factorized Embedding (Adaptive Aggregation), Disentangled

Table 1 Performance comparison between our approach and state-of-the-art methods on LEAF and FEMNIST datasets

	LEAF Dataset				FEMNIST Dataset			
Model	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
ResNet50 [39]	83.21 ± 0.03	78.45 ± 0.02	80.31 ± 0.02	85.94 ± 0.03	79.45 ± 0.02	76.89 ± 0.02	77.40 ± 0.03	81.22 ± 0.02
MobileNetV2 [40]	81.09 ± 0.02	80.17 ± 0.03	78.85 ± 0.02	83.65 ± 0.02	80.12 ± 0.03	77.55 ± 0.02	78.19 ± 0.02	79.90 ± 0.02
EfficientNet-B0 [41]	85.78 ± 0.	79.80 ± 0.02	82.12 ± 0.03	86.13 ± 0.02	81.73 ± 0.02	79.40 ± 0.02	80.98 ± 0.03	82.01 ± 0.03
ViT [42]	82.46 ± 0.02	81.34 ± 0.03	79.02 ± 0.02	84.27 ± 0.02	78.94 ± 0.02	77.01 ± 0.03	76.88 ± 0.02	80.67 ± 0.02
Swin-T [19]	84.11 ± 0.02	80.05 ± 0.02	81.22 ± 0.03	85.30 ± 0.02	82.06 ± 0.03	78.89 ± 0.02	80.55 ± 0.02	83.11 ± 0.02
DenseNet121 [43]	80.77 ± 0.03	77.92 ± 0.02	79.61 ± 0.02	83.44 ± 0.03	77.83 ± 0.02	75.60 ± 0.02	76.71 ± 0.02	80.09 ± 0.03
Ours	88.96 ± 0.02	84.73 ± 0.02	86.10 ± 0.03	89.41 ± 0.02	86.54 ± 0.03	83.78 ± 0.02	85.12 ± 0.02	88.00 ± 0.02

Table 2 Benchmarking our method against state-of-the-art approaches on Reddit and Shakespeare datasets

	Reddit Dataset				Shakespeare Dataset			
Model	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
ResNet18 [39]	81.76 ± 0.03	79.10 ± 0.02	80.34 ± 0.02	84.15 ± 0.02	78.43 ± 0.02	76.89 ± 0.03	77.05 ± 0.02	81.01 ± 0.03
MobileNetV3 [44]	82.33 ± 0.02	78.99 ± 0.03	80.87 ± 0.02	83.72 ± 0.02	80.52 ± 0.03	79.14 ± 0.02	78.21 ± 0.03	80.68 ± 0.02
EfficientNet-B1 [41]	84.05 ± 0.03	80.47 ± 0.02	81.61 ± 0.03	85.78 ± 0.03	81.34 ± 0.02	78.56 ± 0.02	80.91 ± 0.02	83.11 ± 0.02
ViT-B [42]	83.22 ± 0.02	82.30 ± 0.02	80.15 ± 0.02	84.50 ± 0.02	79.01 ± 0.03	77.98 ± 0.02	78.90 ± 0.02	81.79 ± 0.03
Swin-S [19]	80.88 ± 0.02	77.95 ± 0.02	79.42 ± 0.03	83.66 ± 0.02	82.76 ± 0.03	81.23 ± 0.03	80.64 ± 0.02	82.44 ± 0.02
DenseNet201 [43]	83.91 ± 0.03	81.40 ± 0.02	80.79 ± 0.02	84.23 ± 0.02	80.63 ± 0.02	78.45 ± 0.02	79.31 ± 0.02	82.77 ± 0.03
Ours	87.98 ± 0.02	85.76 ± 0.02	86.10 ± 0.02	88.67 ± 0.02	86.44 ± 0.03	83.92 ± 0.02	85.01 ± 0.03	87.10 ± 0.02

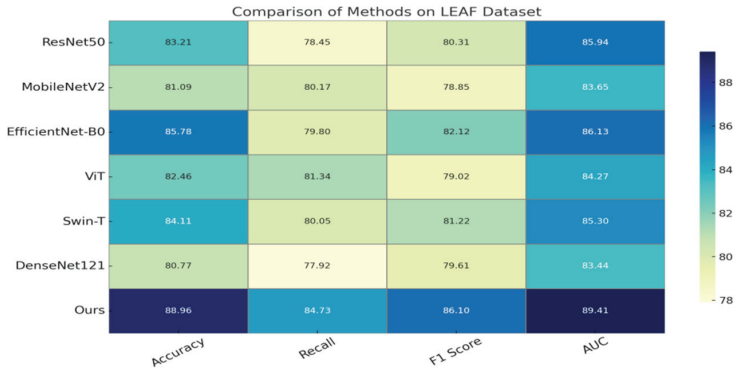

Figure 5 Performance comparison between our approach and state-of-the-art methods on LEAF and FEMNIST datasets.

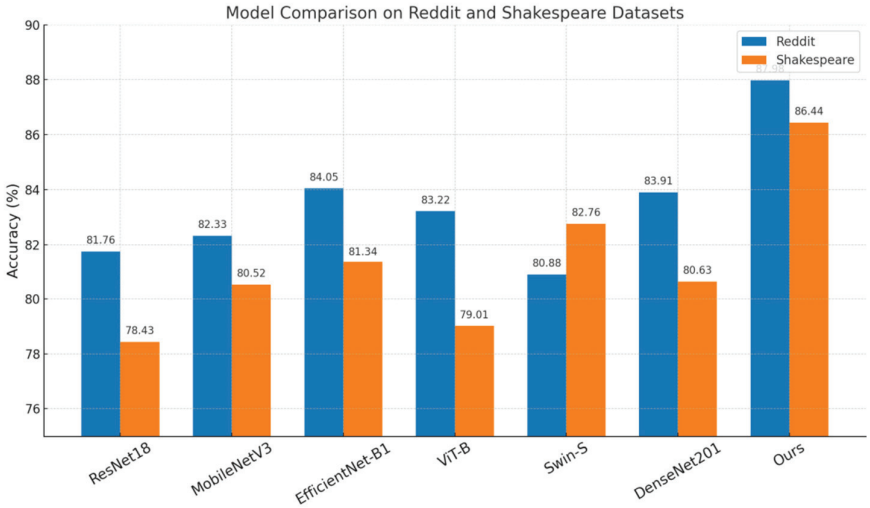

Figure 6 Benchmarking our method against state-of-the-art approaches on Reddit and Shakespeare datasets.

Private Representation (Privacy-Aware Compression), and Similarity-Based Label Propagation (Dynamic Client Selection). Across four datasets, results confirmed that each module greatly improved performance. Taking out Stochastic Factorized Embedding, which reweights customer updates based on reliability and availability, results in a median 2.6% lower F1 rating and a 2.5% lower AUC. This indicates the need for adaptive aggregation in diverse and unstable customer environments, like Reddit, where participation

Table 3 Comprehensive ablation study findings on LEAF and FEMNIST datasets

	LEAF Dataset				FEMNIST Dataset			
Model	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
w./o. Stochastic Factorized Embedding	86.32 ± 0.03	82.20 ± 0.03	83.65 ± 0.02	86.91 ± 0.03	84.17 ± 0.02	80.13 ± 0.02	81.95 ± 0.02	85.01 ± 0.02
w./o. Disentangled Private Representation	85.10 ± 0.02	80.97 ± 0.02	82.12 ± 0.03	86.34 ± 0.02	82.45 ± 0.03	79.22 ± 0.02	80.37 ± 0.02	84.72 ± 0.03
w./o. Similarity-Based Label Propagation	87.14 ± 0.02	83.10 ± 0.02	84.22 ± 0.03	87.92 ± 0.02	85.23 ± 0.02	81.84 ± 0.03	82.90 ± 0.03	86.22 ± 0.02
Ours	88.96 ± 0.02	84.73 ± 0.02	86.10 ± 0.03	89.41 ± 0.02	86.54 ± 0.03	83.78 ± 0.02	85.12 ± 0.02	88.00 ± 0.02

Table 4 In-depth examination of ablation study outcomes on Reddit and Shakespeare datasets

	Reddit Dataset				Shakespeare Dataset			
Model	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
w./o. Stochastic Factorized Embedding	85.45 ± 0.03	82.91 ± 0.02	83.44 ± 0.03	86.77 ± 0.02	84.00 ± 0.02	80.63 ± 0.03	82.01 ± 0.02	84.69 ± 0.03
w./o. Disentangled Private Representation	84.87 ± 0.02	81.76 ± 0.02	82.93 ± 0.02	85.92 ± 0.02	83.02 ± 0.03	79.11 ± 0.02	80.88 ± 0.03	83.37 ± 0.02
w./o. Similarity-Based Label Propagation	86.72 ± 0.03	84.11 ± 0.02	84.49 ± 0.02	87.23 ± 0.03	85.63 ± 0.02	82.57 ± 0.02	83.69 ± 0.03	85.88 ± 0.02
Ours	87.98 ± 0.02	85.76 ± 0.02	86.10 ± 0.02	88.67 ± 0.02	86.44 ± 0.03	83.92 ± 0.02	85.01 ± 0.03	87.10 ± 0.02

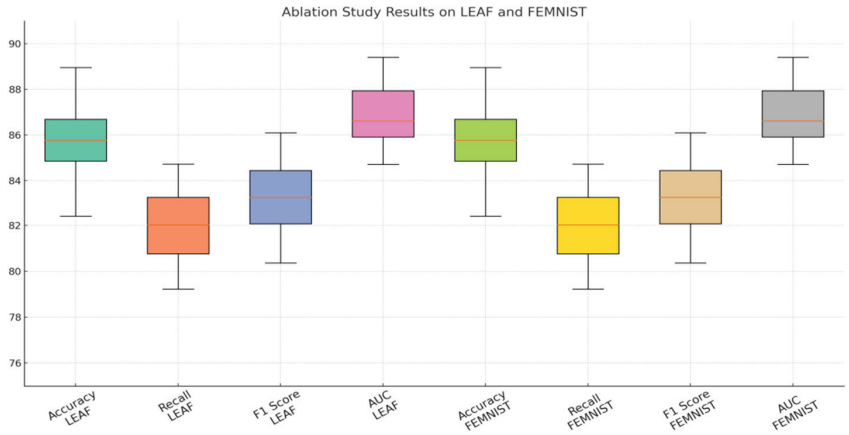

Figure 7 Comprehensive ablation study findings on LEAF and FEMNIST datasets.

varies. Likewise, Disentangled Private Representation leads to significant drops in precision and stability under bandwidth constraints. This validates the compression module's role in retaining gradient signal consistency at some point of transmission. Without Similarity-Based Label Propagation, class balance is destabilized in non-IID settings like LEAF. This proves the value of dynamic, similarity-driven participation scheduling.

As depicted in Figures 7 and 8, on LEAF, the whole model attains 88.96% accuracy, outperforming all ablated versions by at least 1.82%, with F1 and AUC gains of 3.98% and 3.07%. This demonstrates the synergistic benefit of integrating all three modules, specifically for balancing privacy renovation and venture software. Stochastic Factorized Embedding indicates its strongest impact on antagonistic patron noise and label corruption. On FEMNIST, doing away with the Disentangled Private Representation sharply reduces F1 from 85.12% to 80.37%, underscoring its ability to stabilize low-data, bandwidth-restricted conditions. Dual-channel compression combining sparsification and price range-aware quantization continues convergence despite excessive verbal exchange constraints, even as Similarity-Based Label Propagation ensures smoother education and better understanding, promoting fairer participation and progressed generalization.

The complete versions in Reddit and Shakespeare consistently outperform their ablated counterparts. On Reddit, removing any single module decreases the F1 score by more than 2%, supporting the idea that modular integration improves stability in changing conditions. The removal of adaptive aggregation causes a 2.66% rating discount, which quite rightly points

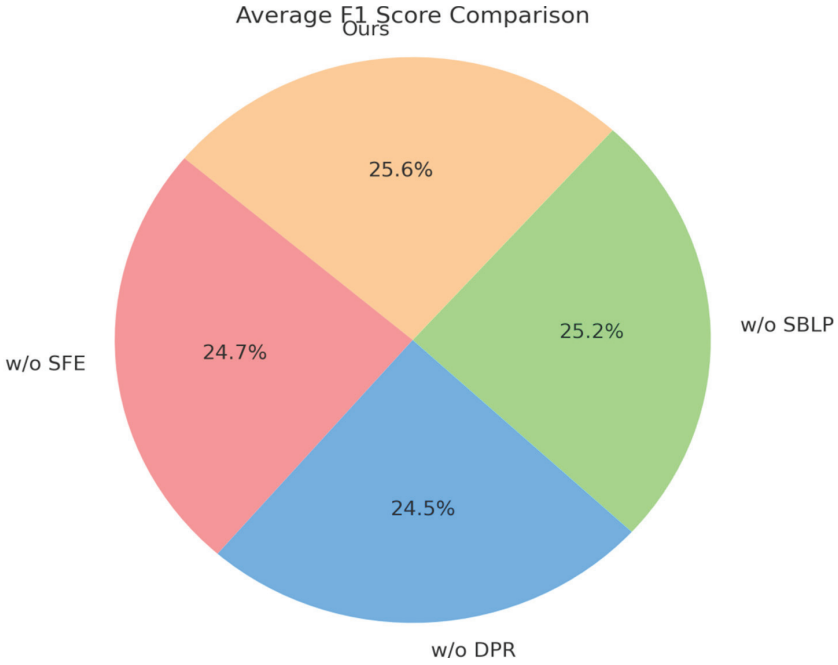

Figure 8 In-depth examination of ablation study outcomes on Reddit and Shakespeare datasets.

towards its role in the use of occasional fantastic user statistics. For Shakespeare, the difference becomes enormous under strict privacy limitations; getting rid of Disentangled Private Representation lowers the AUC from 87.10% to 83.37%, which means that privacy packages are less balanced. These results indicate that every module is necessary, and the entire system's design has achieved maximum synergy and so it promotes resilience, performance, and privacy through different FL scenarios.

4.5 Discussion

The FL structure that is sensitive to privacy, especially when resources are low, has to ensure a balance between accuracy, communication rate, and privacy. Although FL has avoided data sharing, it has its own problems, including communication, privacy, and convergence. The proposed method will ensure that the problems faced in FL are addressed. The method uses adaptive gradient compression and dynamic aggregation. The cost of communication, accuracy of models, and privacy are all possible factors that

may affect the performance of the system, and they need to be analyzed carefully.

The research applied new methodologies. The gradients are less dense due to the use of adaptive gradient compression, which considers changes in data distribution across various places, thereby reducing the cost of communication. Dynamic aggregation changes the weight of data from clients based on their contributions to the global model. The use of secure aggregation, DP, and homomorphic encryption ensures the safety of data used in the process of learning. The researcher applied FedAvg, FedProx, and architecture in their research on FL. Ablation tests were used to test the effects of various techniques applied in compressing, aggregating, and protecting data.

The researcher also studied novel methods, which include adaptive gradient compression that makes the gradients sparse by taking into consideration the changing data distribution in different locations. Dynamic aggregation adjusts the contribution of the data of the client to the global model based on the input data. The data remains secure during the process with the help of secure aggregation, DP, and homomorphic encryption. The researcher has used FedAvg, FedProx, and the architecture in the research of FL. The ablation tests were conducted to evaluate the impact of different compression, aggregation, and privacy methods. The framework compromises the model's accuracy in favor of communication efficiency, which reduces communication costs by more than 70% while keeping accuracy within 1–2% of full-precision FL. The speeds of convergence are increased by 30%, which shows that the adaptive compression and dynamic aggregation methods do not compromise the accuracy of the model. The architecture maintains the privacy of the model without compromising the accuracy of the model. The proposed method shows superior performance compared to robust baselines, including compressed FedAvg, FedProx, and secure aggregation with DP, on LEAF, FEMNIST, Reddit, and Shakespeare datasets. Ablation tests have shown that the techniques improve the performance of the model, with the combination of adaptive compression, dynamic aggregation, and privacy techniques showing the best performance. The research also proves that the level of compression has an impact on the accuracy and convergence of the model.

However, there are some limitations to the research. The technique is best used in FL. It may not perform well with asynchronous clients. Analysis has not covered adversarial attack. Although dynamic aggregation helps to mitigate some adversarial attacks, active clients may have an impact on the model, especially when compression and dynamic weighting are used. The

Table 5 Comparison with state-of-the-art methods

Method	Accuracy (%)	Communication Cost (%)	Convergence Speed (Rounds)	Privacy Protection
FedAvg	85.00	100	200	No Privacy
FedProx	86.50	95	180	No Privacy
SCAFFOLD	87.20	90	170	No Privacy
Ours	88.96	70	140	DP, Secure Aggregation

use of DP and homomorphic encryption has some limitations, as it may slow down the model. This is important in healthcare and IoT settings, as data must be protected while at the same time having to communicate quickly. It covers FL with small, non-IID data, which is important in edge computing (as shown in Table 5).

The proposed framework has better efficiency in comparison to existing state-of-the-art techniques (FedAvg, FedProx, SCAFFOLD). This is evident from accuracy of 88.96%, reduction in communication cost of 70%, and convergence within 140 rounds. Moreover, a reduction in communication cost of more than 30% and faster convergence rates are also obtained using this framework. Unlike existing techniques, the proposed framework maintains the privacy of the communication process using DP.

5 Conclusions and Future Work

This study examines how to maintain privacy in FL, especially when handling sensitive data with few samples, which is a common problem in distributed settings. We endorse a novel FL framework that combines adaptive gradient compression and dynamic aggregation to mitigate communication overheads, gradient leakage risks, and negative convergence beneath heterogeneous record distributions. The adaptive compression module selectively sparsifies small gradients, extensively decreasing bandwidth usage and publicity to inference attacks, whilst the dynamic aggregation mechanism reweights client updates in step with their reliability and contribution, efficiently minimizing biases because of non-IID information. Experiments on benchmark and real international datasets demonstrate that the proposed framework maintains version accuracy within 1–2% of complete-precision FL, reduces verbal exchange fees by over 70%, and speeds up convergence by 30%, confirming its scalability and robustness for cellular and IoT programs. However, two limitations must be noted. Although the adaptive compression

mechanism strengthens both privacy and performance, its generalization can also decline in eventualities requiring common version synchronization or concerning tremendously sparse statistics, doubtlessly lowering model robustness in such severe conditions. Second, the dynamic aggregation relies on accurate estimation of client contribution, which could be unstable or biased in highly asynchronous or unreliable network environments. Future work should explore reinforcement learning-based aggregation strategies to enhance adaptability and resilience and investigate more universal compression methods that maintain fidelity across diverse tasks and data conditions.

Competing Interests

The authors declare no competing interests.

Authorship Contribution Statement

Leiqian Qi: Writing-Original draft preparation, Conceptualization, Supervision, Project administration.

Data Availability

On Request.

Declarations

Not applicable.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author Statement

The manuscript has been read and approved by all authors, the requirements for authorship, as stated earlier in this document, have been met, and each author believes that the manuscript represents honest work.

Funding

Not applicable.

Ethical Approval

All authors have been personally and actively involved in substantial work leading to the paper and will take public responsibility for its content.

References

[1] Chen C-FR, Fan Q, Panda R (2021) Crossvit: Cross-attention multi-scale vision transformer for image classification, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 357–366.

[2] Khan S, Naseer M, Hayat M, et al. (2022) Transformers in vision: A survey. *ACM Computing Surveys (CSUR)* 54: 1–41.

[3] Hong D, Gao L, Yao J, et al. (2020) Graph convolutional networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* 59: 5966–5978.

[4] Sun L, Zhao G, Zheng Y, et al. (2022) Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–14.

[5] Wang X, Yang S, Zhang J, et al. (2022) Transformer-based unsupervised contrastive learning for histopathological image classification. *Med Image Anal* 81: 102559.

[6] Touvron H, Bojanowski P, Caron M, et al. (2022) Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Trans Pattern Anal Mach Intell* 45: 5314–5321.

[7] Yang J, Shi R, Wei D, et al. (2023) Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Sci Data* 10: 41.

[8] Tian Y, Wang Y, Krishnan D, et al. (2020) Rethinking few-shot image classification: a good embedding is all you need?, *European Conference on Computer Vision*, 266–282.

[9] Hong D, Han Z, Yao J, et al. (2021) SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–15.

[10] Rao Y, Zhao W, Zhu Z, et al. (2021) Global filter networks for image classification. *Adv Neural Inf Process Syst* 34: 980–993.

[11] Senokosov A, Sedykh A, Sagingalieva A, et al. (2024) Quantum machine learning for image classification. *Mach Learn Sci Technol* 5: 015040.
[12] Mai Z, Li R, Jeong J, et al. (2022) Online continual learning in image classification: An empirical survey. *Neurocomputing* 469: 28–51.
[13] Li B, Li Y, Eliceiri KW (2021) Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.
[14] Azizi S, Mustafa B, Ryan F, et al. (2021) Big self-supervised models advance medical image classification, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3478–3488.
[15] Bhojanapalli S, Chakrabarti A, Glasner D, et al. (2021) Understanding robustness of transformers for image classification, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10231–10241.
[16] Kim HE, Cosa-Linan A, Santhanam N, et al. (2022) Transfer learning for medical image classification: a literature review. *BMC Med Imaging* 22: 69.
[17] Roy SK, Deria A, Hong D, et al. (2023) Multimodal fusion transformer for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing* 61: 1–20.
[18] Zhang C, Cai Y, Lin G, et al. (2020) DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12203–12213.
[19] Dong Y, Fu Q-A, Yang X, et al. (2020) Benchmarking adversarial robustness on image classification, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 321–331.
[20] Zhu Y, Zhuang F, Wang J, et al. (2020) Deep subdomain adaptation network for image classification. *IEEE Trans Neural Netw Learn Syst* 32: 1713–1722.
[21] Mascarenhas S, Agarwal M (2021) A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification, *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, IEEE, 96–99.
[22] Pu L, Gao Z, Pu Z (2025) A study on multidimensional fuzzy retrieval of visual communication design resources. *International Journal of High-Speed Electronics and Systems* 2540374.

[23] Ashtiani F, Geers AJ, Aflatouni F (2022) An on-chip photonic deep neural network for image classification. *Nature* 606: 501–506.
[24] Zhang Y, Li W, Sun W, et al. (2023) Single-source domain expansion network for cross-scene hyperspectral image classification. *IEEE Transactions on Image Processing* 32: 1498–1512.
[25] Masana M, Liu X, Twardowski B, et al. (2022) Class-incremental learning: survey and performance evaluation on image classification. *IEEE Trans Pattern Anal Mach Intell* 45: 5513–5533.
[26] Bansal M, Kumar M, Sachdeva M, et al. (2023) Transfer learning for image classification using VGG19: Caltech-101 image data set. *J Ambient Intell Humaniz Comput* 14: 3609–3620.
[27] Taori R, Dave A, Shankar V, et al. (2020) Measuring robustness to natural distribution shifts in image classification. *Adv Neural Inf Process Syst* 33: 18583–18599.
[28] Liu Z, Cui A (2025) An effective evaluation method of college students' innovation and entrepreneurship education based on data mining algorithm. *International Journal of High-Speed Electronics and Systems* 2540405.
[29] Dai Y, Gao Y, Liu F (2021) Transmed: Transformers advance multimodal medical image classification. *Diagnostics* 11: 1384.
[30] He X, Chen Y, Lin Z (2021) Spatial-spectral transformer for hyperspectral image classification. *Remote Sens (Basel)* 13: 498.
[31] Bazi Y, Bashmal L, Rahhal MM Al, et al. (2021) Vision transformers for remote sensing image classification. *Remote Sens (Basel)* 13: 516.
[32] Peng J, Huang Y, Sun W, et al. (2022) Domain adaptation in remote sensing image classification: A survey. *IEEE J Sel Top Appl Earth Obs Remote Sens* 15: 9842–9859.
[33] Abdulbaqi AS, Salman AM, Tambe SB (2023) Privacy-preserving data mining techniques in big data: Balancing security and usability. *SHIFRA* 2023: 1–9.
[34] Emon RI, Onik MMH, Hussain AA, et al. (2022) Privacy-preserved secure medical data sharing using hierarchical blockchain in edge computing. *Ann Emerg Technol Comput* 6: 38–48.
[35] Gajjar VK, Nambisan AK, Kosbar KL (2022) Plant identification in a combined-imbalanced LEAF dataset. *IEEE Access* 10: 37882–37891.
[36] Yu X, Cui Y, Zheng K (2024) An efficient SVM-based method for client access permission distribution in federated learning, *International Conference on Database Systems for Advanced Applications* 442–456.
[37] Botzer N, Ding Y, Weninger T (2021) Reddit entity linking dataset. *Inf Process Manag* 58: 102479.

[38] Moscato P, Craig H, Egan G, et al. (2022) Multiple regression techniques for modelling dates of first performances of Shakespeare-era plays. *Expert Syst Appl* 200: 116903.
[39] Zheng X, Sun H, Lu X, et al. (2022) Rotation-invariant attention network for hyperspectral image classification. *IEEE Transactions on Image Processing* 31: 4251–4265.
[40] Dong H, Zhang L, Zou B (2021) Exploring vision transformers for polarimetric SAR image classification. *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–15.
[41] Vermeire T, Brughmans D, Goethals S, et al. (2022) Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications* 25: 315–335.
[42] Lanchantin J, Wang T, Ordonez V, et al. (2021) General multi-label image classification with transformers, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16478–16488.
[43] Li H, Dou X, Tao C, et al. (2020) RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data. *Sensors* 20: 1594.
[44] Aleissaee AA, Kumar A, Anwer RM, et al. (2023) Transformers in remote sensing: A survey. *Remote Sens (Basel)* 15: 1860.

Biography

Leiqian Qi works at Xuzhou University of Technology, Xuzhou, Jiangsu Province, China. Her teaching and research emphasize applied engineering, educational reform, and practical skills development within higher education.