Adaptive Confidence-Driven Sample Selection with Alignment Loss for Efficient Federated Active Learning

Huiping Zhou *

School of Computer Science, Nanjing University, Nanjing, China

* Corresponding Author Email: 221220039@smail.nju.edu.cn

Abstract. Federated Active Learning (FAL) aims to reduce labeling costs by selecting the most informative samples for annotation while preserving strong model performance. However, current FAL methods struggle with the challenge of client heterogeneity in non-IID environments, often leading to suboptimal sample selection and increased computational burden due to repeated inference on unlabeled data. To address these issues, we propose a novel approach called Adaptive Confidence-Driven Sample Selection (ACdSS), which combines the CHASe strategy with Alignment Loss to improve sample selection efficiency. ACdSS leverages dynamic confidence scores to assess uncertainty in unlabeled data and identifies samples with fluctuating predictions between training rounds, which are more likely to be near the decision boundary. This approach balances exploration and exploitation by adapting the sample selection strategy throughout the training process, minimizing inference costs while maintaining high selection effectiveness. Experimental results on benchmark datasets demonstrate that ACdSS outperforms existing FAL techniques, delivering superior accuracy with reduced annotation costs. This framework offers a practical solution for managing client heterogeneity in FAL and is well-suited for real-world federated learning systems, offering a significant reduction in computational overhead while improving model performance.

Keywords: Machine Learning, Label-Flipping Attack, Federated Learning, Client Weighting.

1. Introduction

Federated Active Learning (FAL) combines the advantages of federated learning (FL) and active learning, aiming to reduce the cost of labeling data while maintaining robust model performance in decentralized environments. FAL is particularly appealing in situations where labeled data is scarce or expensive to obtain, such as medical imaging or natural language processing, as it prioritizes the selection of the most informative samples for annotation. However, one of the most significant challenges faced by FAL is handling client heterogeneity, particularly in non-IID (Independent and Identically Distributed) settings, where data distributions across clients vary significantly. This issue leads to suboptimal performance in sample selection, as current methods tend to overlook the temporal evolution of uncertainty in model predictions.

Traditional FAL strategies that monitor prediction uncertainty often require repeated inferences on unlabeled data, which can result in substantial computational overhead, especially in federated settings with limited resources. Moreover, these techniques may not effectively handle the dynamic and evolving nature of sample uncertainty across different stages of the training process. While confidence-based or margin-based approaches are computationally efficient, they fail to account for the changes in prediction behavior over time, missing valuable patterns that can help improve the model's generalization.

To overcome these limitations, we propose a novel framework, Adaptive Confidence-Driven Sample Selection (ACdSS), which addresses both the challenges of client heterogeneity and the high computational burden associated with sample selection in federated active learning. ACdSS builds upon the CHASe strategy for client heterogeneity awareness and incorporates Alignment Loss for stability in decision boundaries across communication rounds. By utilizing confidence scores and dynamically adjusting the selection strategy, ACdSS identifies the most informative samples while maintaining efficiency in resource-constrained federated environments. This approach balances exploration and exploitation, offering an adaptive selection mechanism that evolves throughout the training process [1,2].

The key innovations of ACdSS lie in its ability to leverage prediction inconsistencies between rounds, dynamic uncertainty estimation, and alignment loss to stabilize boundary-sensitive samples. These techniques ensure that the sample selection process is both computationally efficient and effective, addressing the unique challenges posed by non-IID data distributions and client heterogeneity. Through extensive experiments on benchmark datasets, our approach demonstrates superior performance compared to existing FAL strategies, offering a scalable and efficient solution for real-world federated learning systems [3].

2. Relevant Theories

2.1. Federated Learning

Federated learning is a decentralized machine learning framework in which a group of distributed clients work together to train a shared global model without exchanging their local data. This setup enhances data privacy and minimizes the need to transfer raw data across networks. Widely used algorithms like FedAvg allow each client to conduct multiple rounds of local training before sending their model updates to a central server for aggregation [4, 5]. One of the primary difficulties in federated learning is managing data heterogeneity across clients, commonly referred to as non-IID data distribution, which can significantly impact the performance of the global model if not properly handled.

In federated learning, clients typically operate on locally collected data that may differ significantly in terms of distribution, volume, and quality. This divergence introduces challenges in ensuring that the global model performs consistently across all participating clients. Additionally, the communication constraints between clients and the central server must be considered, as frequent synchronization can lead to high latency and increased bandwidth usage. To address these issues, various optimization techniques have been proposed, including adaptive aggregation strategies, personalized model updates, and compression methods for model transmissions. These approaches aim to improve convergence rates and overall model accuracy while maintaining the benefits of decentralized data processing.

2.2. Active Learning

Active learning is a machine learning approach designed to reduce labeling effort by strategically selecting the most informative unlabeled samples for annotation. In the pool-based setting, the learning algorithm evaluates the entire pool of unlabeled data and selects instances where the model exhibits the highest uncertainty—often quantified using metrics such as entropy, margin, or uncertainty measures [6]. This iterative process allows the model to improve over time while requiring only a small fraction of the dataset to be labeled.

The core idea behind active learning is to prioritize data points that are expected to yield the greatest performance gain when labeled. This is particularly valuable in domains where labeling is expensive, time-consuming, or requires expert knowledge, such as medical imaging or natural language processing. Various query strategies have been developed to identify informative samples, including uncertainty sampling, query-by-committee, and expected model change maximization. These strategies aim to balance exploration and exploitation, ensuring that the selected samples contribute meaningfully to the model's learning process. However, active learning also faces challenges, such as the risk of selecting noisy or unrepresentative samples, which can hinder model performance if not carefully managed.

2.3. Federated Active Learning Approaches

Federated active learning (FAL) integrates the concepts of federated learning and active learning to enable efficient and privacy-preserving label acquisition. It must simultaneously manage data heterogeneity across clients and limitations on labeling resources. Current FAL techniques include methods based on uncertainty sampling, margin-driven selection, diversity preservation, and

measuring the discrepancy between local and global model knowledge [7, 8]. These approaches aim to identify data samples that are both locally uncertain and globally informative. However, many existing strategies either introduce high computational demands—such as repeated model inference—or neglect the temporal evolution of sample uncertainty during training.

In FAL, the challenge lies in coordinating the selection of informative samples across distributed clients while preserving data privacy and minimizing communication overhead. Some approaches propose that clients perform local uncertainty estimation and send only the most informative samples for labeling, either by the server or designated annotators. Others incorporate diversity-aware selection mechanisms to ensure that the chosen samples represent the overall data distribution across clients. Additionally, recent studies have explored dynamic strategies that adaptively adjust the selection criteria based on the model's learning progress. Despite these advancements, there remains a need for scalable and computationally efficient methods that can effectively handle evolving data distributions and varying client participation patterns in real-world deployments.

3. Problem Definition

Federated Active Learning (FAL) integrates the principles of federated learning and active learning with the objective of reducing labeling costs while ensuring data privacy. In this setup, a central server coordinates the learning process across multiple distributed clients, where each client maintains a private, non-IID collection of unlabeled data. During training, each client is responsible for selecting a limited number of unlabeled samples for annotation under a predefined annotation budget. After obtaining the corresponding labels, the client trains its local model and transmits model updates to the central server to contribute to the global model refinement.

Let K represent the total number of clients, and for each client k, denote U_k as its current pool of unlabeled data and L_k as its labeled dataset. In each global round t, client k selects a subset $Q_k^t \subset U_k$ of size B (where B is the annotation budget). It then receives the true labels for Q_k^t, updates its labeled set as $L_k \leftarrow L_k \cup Q_k^t$, proceeds with local model training using L_k, and uploads the resulting model updates to the server [9].

The key challenge lies in developing an effective sample selection criterion $s(x)$, where $x \in U_k$, that measures the informativeness of unlabeled instances. The criterion must strike a balance between selecting samples with high uncertainty and maintaining low computational overhead. Specifically, the selection strategy should satisfy the following requirements [10]:

• It should be efficiently computable on each client with minimal inference cost per communication round between the server and clients.

• It should be robust to non-IID data distributions across clients, supporting convergence of the global model.

• It should incorporate temporal dynamics or sensitivity to the decision boundary without necessitating inference at every training epoch.

To address these challenges, the proposed ACdSS approach introduces a selection mechanism that evaluates both the variation in predictions across successive global rounds and the current prediction entropy (a measure of uncertainty). This results in a computationally efficient yet effective strategy for identifying the most valuable samples for annotation.

Prediction inconsistency detection: identifying boundary-sensitive samples: To prioritize the selection of informative instances, we focus exclusively on unlabeled samples that exhibit label prediction shifts between successive communication rounds $t-1$ and t. Such samples are typically located in proximity to decision boundaries, where prediction uncertainty is highest, making them particularly valuable for model improvement. Previous studies have shown that instances displaying inconsistent predictions across training rounds tend to reside in critical regions near classification boundaries and often provide substantial learning signals when labeled.

Entropy computation and uncertainty estimation, for each candidate sample i, compute its prediction entropy at round t:

$$H(p_i^t) = -\sum_{c=1}^{C} p_{i,c}^t \log p_{i,c}^t \quad (1)$$

And record the entropy difference across rounds $\left|H(p_i^t) - H(p_i^{t-1})\right|$, which captures the model's confidence evolution over time.

Dynamic uncertainty score, Define a combined score for each sample:

$$U(i) = H(p_i^t) + \lambda_t \cdot \left|H(p_i^t) - H(p_i^{t-1})\right|$$
$$where\ \lambda_t = \lambda_0 \exp(-k \cdot \max(0, t-25)) \quad (2)$$

Decays over training rounds, emphasizing confidence fluctuation initially and current entropy later. Generally speaking, entropy is used to calculate the amount of information contained in a sample. FedEntropy directly uses entropy as the basis for sorting. However, ACdSS uses entropy to calculate confidence. That is, samples with high entropy as perceived by the model do not necessarily contain the information needed by the local model. They could be information that the model is relatively certain about or errors in the model's results. By introducing the entropy calculated by the previous model, samples that the model is less certain about can be filtered out. Especially as the number of rounds increases and the model's accuracy gradually improves, the difference in the amount of information judged by the model in two consecutive rounds becomes less important. In summary, the sample selection strategy of ACdSS is to improve selection efficiency by choosing samples that are based on two rounds of model judgments, have high credibility, and possess both uncertain results and high information content as determined by the model.

Adaptive random–confidence sampling

Let labeling budget be B, define sampling ratio α_t:

$$\alpha_t = \begin{cases} 0, & t \le 25, \\ \min\left(1, \frac{t-25}{T_{max}-25}\right), & t > 25. \end{cases} \quad (3)$$

Select $\alpha_t B$ samples with highest $U(i)$, and randomly choose the remaining $(1-\alpha_t)B$ from other candidates. This approach endeavors to strike a balance between reducing the computational burden (by randomly selecting and omitting the inference calculations of a portion of the data) and attaining a better sample selection outcome (computed based on the amount of information). Logically, when a model is just beginning its training, its accuracy is usually low, and there are significant errors in the entropy/confidence of the inference results [11]. To a certain extent, randomly selected samples can be used as substitutes, and similar results can still be maintained.

Design rationale and efficiency comparison: The algorithm only performs one inference per communication round, avoiding repeated epoch-level inference. This design achieves computational efficiency comparable to standard FedMargin or entropy-based methods, while still capitalizing on boundary unstable samples valuable for active selection [12].

3.1. Alignment Loss for Boundary Stability

An Alignment Loss term is introduced to enhance boundary stability by reducing discrepancies between local and global model representations on boundary-sensitive samples. This module helps mitigate local model drift caused by heterogeneous data distributions across clients.

Implementation steps:

• After local training in communication round t, extract representations for samples in the candidate set C using the current local model ω_k^t, global model ω_{global}^t, and previous-round local model ω_k^{t-1}.

• Compute the alignment loss using a model-level contrastive objective, for example:

$$l_{\text{align}}(x_i) = -\log \frac{\exp\left(\text{sim}\left(z_i^k, z_i^{\text{global}}\right)/\tau\right)}{\exp\left(\text{sim}\left(z_i^k, z_i^{\text{global}}\right)/\tau\right) + \exp\left(\text{sim}\left(z_i^k, z_i^{\text{prev}}\right)/\tau\right)} \quad (4)$$

Here, z_i^k, z_i^{global}, and z_i^{prev} are feature vectors from current local model, global model, and previous-round local model, respectively; sim denotes cosine similarity, and τ is a temperature hyperparameter.

This formulation aligns closely with the MOON framework (Model-Contrastive Federated Learning), which uses a contrastive regularization term to make local model representations approach the global model and diverge from the previous local model, improving robustness in non-IID federated settings [13].

In contrast to MOON, which employs a model-level contrastive loss to align the local model with the global model while distancing it from its prior state during each local training round, ACdSS introduces a more focused and computationally efficient Alignment Loss mechanism:

- Alignment loss is computed exclusively for part of the unlabeled instances—significantly lowering the inference cost.
- The alignment process considers both the global model and the local model from the previous round, aiming to refine and stabilize the decision boundaries.
- Unlike MOON, ACdSS preserves the integrity of the standard federated learning pipeline. The alignment loss serves as an auxiliary regularization term, applied solely during the local training phase.

This approach retains the advantage of contrastive alignment in reducing client drift, without incurring the computational cost of full-epoch contrastive training. It effectively balances model performance and computational efficiency, rendering ACdSS particularly well-suited for federated active learning settings with constrained resources.

4. Experiments and Evaluation

The empirical evaluation follows standard federated active learning benchmarks to assess the effectiveness and efficiency of ACdSS in comparison with baseline strategies such as FALMargin, FALEntropy, and FALRandom.

4.1. Datasets and Partitioning

Experiments are carried out on the widely adopted image classification benchmark, CIFAR-10. To emulate heterogeneous (non-IID) data distributions among clients, the training data is divided across clients using Dirichlet distributions with different concentration levels (e.g., $\alpha \in \{0.1,0.3,0.5,0.8,1.0\}$) [14,15].

4.2. Client Participation and Budget

Evaluation Metrics: Model effectiveness is evaluated based on test set accuracy over a set number of communication rounds. Efficiency is analyzed in terms of execution time and inference overhead per FAL round.

Federated Learning Configuration: the total number of clients is set to $K = 10$. Each client holds data from $C_k = 5$ distinct classes, simulating non-IID conditions. Under the assumption of honest and non-malicious clients [13], we consider two client collaboration modes: In the Absolute Cooperation scenario, all clients participate fully with identical annotation budgets; in the Relative Cooperation setting, annotation budgets differ according to each client's level of involvement.

4.3. Model Architecture and Training

• A simple convolutional neural network (simple-CNN) is used to accommodate fair comparison across baselines.

• Local training consists of E local epochs per client, using standard cross-entropy loss and SGD optimizer with a fixed learning rate.

4.4. Baselines and Metrics

• Baselines include FALRandom (random sampling), FALEntropy (sampling based on Entropy), FedMargin (sampling based on Margin).

• Evaluation metrics: global test accuracy, performance curve, and variance across sampling strategies.

Hardware and Implementation

• All methods are implemented in PyTorch, experiments run on GPU. Results are averaged over multiple random seeds to ensure robustness.

5. Results

Figure 1 presents a comparison of test accuracy among various active sampling techniques throughout the federated learning process. The x-axis illustrates communication rounds ranging from 0 to 200, while the y-axis reflects the corresponding test accuracy. The evaluated methods include ACdSS (blue), chase-subset500 (green), FALEntropy (red), and FALMargin (purple).

In the initial phase (rounds 0–50), all methods experience a sharp rise in accuracy, which is consistent with expectations. Among them, chase-subset500 achieves the fastest growth and sustains the highest performance throughout training, reaching a peak accuracy slightly above 0.43. This outcome indicates that chase-subset500 benefits from its strategy of identifying boundary-aware and fluctuation-sensitive samples, thereby selecting more informative data points. ACdSS, which incorporates both adaptive confidence estimation and inconsistency monitoring, demonstrates competitive performance. It frequently outperforms FALMargin and FALEntropy after round 50. Importantly, ACdSS maintains a steady upward trend in accuracy with minimal fluctuations compared to FALMargin, suggesting superior stability and more effective sample selection over time. FALEntropy and FALMargin deliver promising initial results but eventually converge to a lower performance plateau, approximately 0.4. This implies that although these methods are computationally efficient, they may not capture the informativeness of samples as effectively as ACdSS and chase-subset500. In summary, the figure illustrates the trade-off between computational efficiency and model performance. ACdSS strikes a favorable balance—achieving performance close to that of chase-subset500 while preserving the efficiency typical of simpler heuristic-based approaches.

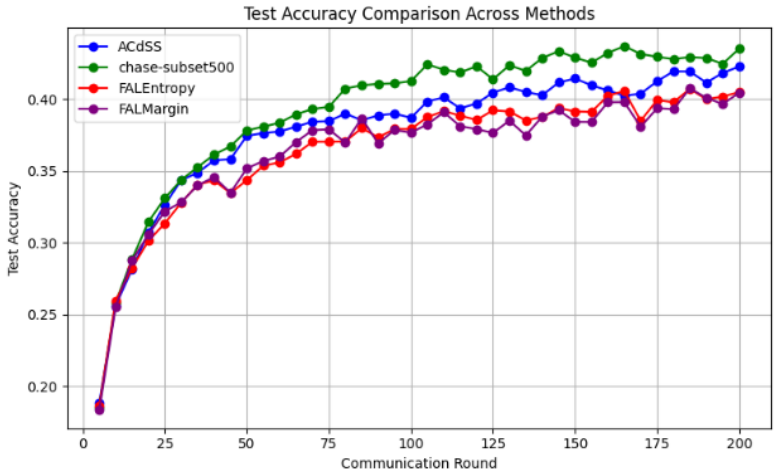


Figure 1. The training loss dynamics.

Figure 1 presents a comparison of the training loss dynamics among various active learning approaches within the federated learning framework.

As training progresses, all methods show a gradual reduction in training loss. Interestingly, ACdSS and chase-subset500 start with higher loss values compared to FALEntropy and FALMargin. This can be largely attributed to their incorporation of alignment loss, which imposes extra constraints

during training to enhance consistency between local and global model representations. Although this alignment mechanism guides the learning process more effectively, it naturally elevates the overall loss in the early stages. As training continues, FALMargin attains the lowest training loss, which is anticipated since it relies purely on cross-entropy loss without any additional regularization components. The training loss of ACdSS remains slightly higher than that of FALMargin, yet it exhibits a more stable and consistent decline, showing the most progressive improvement over time and indicating a smoother optimization landscape. These findings imply that while ACdSS may not yield the minimal training loss, it achieves a favorable balance by utilizing alignment-based regularization to select more informative samples, even at the cost of a slight increase in loss magnitude.

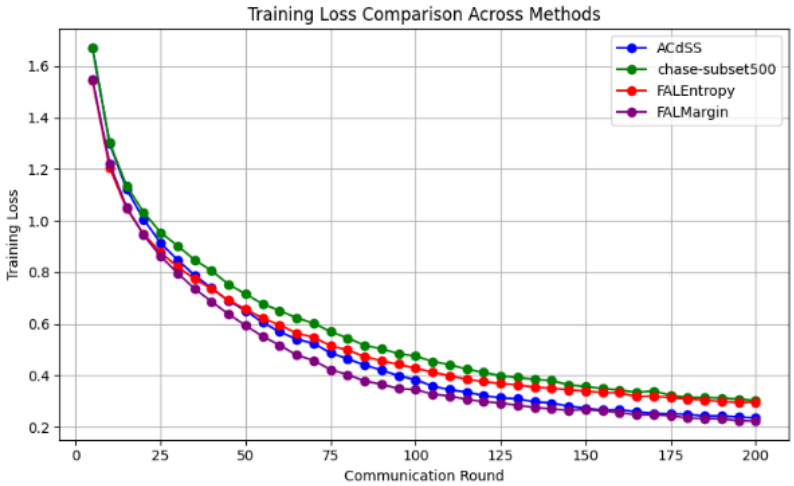


Figure 2. The training accuracy trajectories.

Figure 2 illustrates the training accuracy trajectories of four federated active learning strategies over the course of 200 communication rounds.

As the training proceeds, all methods exhibit a gradual and consistent rise in training accuracy. Among them, chase-subset500 consistently achieves the highest accuracy across the majority of rounds. This outcome is anticipated, as chase-subset500 employs a more aggressive boundary-aware selection strategy combined with larger data subsets, which enhances the quality of local model training.

ACdSS demonstrates strong performance, ranking second in terms of training accuracy after chase-subset500. The steady improvement in its accuracy supports the effectiveness of its hybrid sample selection approach, which effectively balances exploration through randomness and exploitation based on model confidence. Compared to FALEntropy and FALMargin, ACdSS shows better results in the later stages of training, highlighting its ability to progressively focus on informative boundary samples.

In contrast, FALMargin displays greater variability and lower overall accuracy, potentially due to its dependence on a static margin-based uncertainty measure, which may not adequately adapt to the changing decision boundaries during training. Meanwhile, it is noted that the advantage of ACdSS in training accuracy (Acc) relative to the baseline is not as pronounced as it is in testing. This may indicate that ACdSS offers certain insights and advantages in the training of aggregated models, namely non-specialized models.

The Figure 3 illustrates the training accuracy comparison across different federated active learning methods over 200 communication rounds. The x-axis represents the number of communication rounds, while the y-axis shows the training accuracy achieved by each method. Among the methods evaluated, ACdSS (blue line) consistently outperforms the others, achieving the highest training accuracy and maintaining steady improvement throughout the training process. This suggests that the Adaptive Confidence-Driven Sample Selection approach is highly effective in identifying and prioritizing informative samples, which enhances model performance. Chase Subset500 (green line)

also demonstrates strong performance, with a steady increase in accuracy, though it levels off slightly before reaching the peak accuracy achieved by ACdSS. In contrast, MultiKMeans (red line) and FedAvg (purple line) exhibit slower improvements and achieve lower overall accuracy, highlighting their relatively less effective sample selection strategies in the presence of client heterogeneity. These results underscore the superior stability and effectiveness of ACdSS in optimizing training accuracy in federated active learning scenarios.

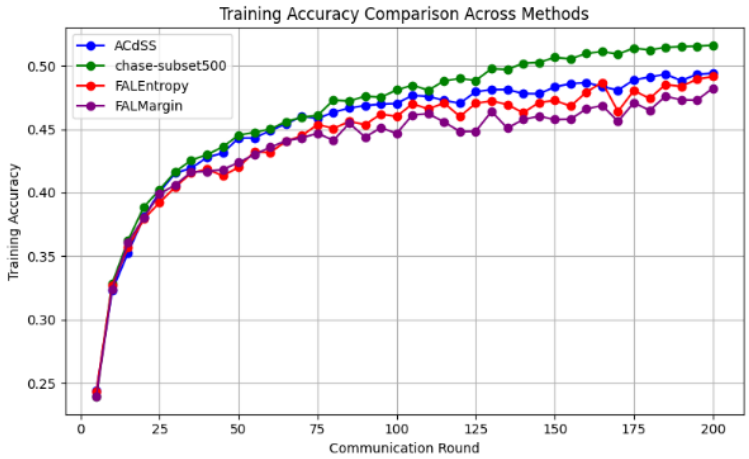


Figure 3. Training Accuracy Comparison Across Methods.

6. Challenges and Future Work

Although ACdSS demonstrates strong performance and computational efficiency in federated active learning, several challenges remain: Robustness under highly skewed data distributions: In highly non-IID settings, where significant heterogeneity exists among client data distributions, the existing selection approach may face challenges in preserving coherence and ensuring effective generalization across different clients. Limited label budget and noisy annotations: The approach relies on the assumption of precise human annotations during querying; however, in practical settings, labeling resources are typically constrained, and annotated data may contain inaccuracies or noise, both of which can negatively impact model performance. Scalability to large models and datasets: Although ACdSS minimizes the frequency of inference operations, it still involves the storage and comparison of predictions across multiple training rounds. In large-scale scenarios featuring high-dimensional data or complex models, this requirement may introduce significant memory and computational overhead, potentially becoming a limiting factor in system scalability.

To further enhance ACdSS and broaden its applicability, several research directions can be pursued: Adaptive selection in client-variant environments: Implement adaptive selection approaches tailored to individual clients, which modify the selection criteria in response to local data characteristics or ongoing training performance.

Robustness-aware uncertainty estimation: Investigate methods for quantifying uncertainty that remain effective in the presence of noisy labels or adversarial examples, thereby enhancing sample selection reliability in imperfect data conditions.

Integration with scalable inference and communication strategies: Incorporate strategies such as model compression, gradient sparsification, or knowledge distillation alongside ACdSS to minimize both communication overhead and computational demands.

7. Conclusion

This study presents a lightweight and privacy-preserving defense method for federated learning under label-flipping attacks. By integrating client-side validation accuracy and update consistency

into the aggregation process, the proposed strategy seeks to reduce the influence of poisoned updates without relying on raw data or introducing excessive computation.

Experimental evaluations on the Fashion-MNIST dataset under three distinct attack Configurations —(1 →6), (2 →7), and (4 → 9)—indicate that the method achieves slightly higher accuracy and improved training stability compared to FedAvg and Trimmed Mean. However, the improvements over FedAvg were relatively minor in most cases, highlighting both the potential and the limitations of relying solely on statistical signals like validation performance and update drift for detecting adversarial behavior.

Despite the modest gains, the method demonstrates practical advantages: it is efficient, compatible with existing FL pipelines, and requires no prior assumptions about the number or identity of malicious clients. These traits make it a suitable defense option in scenarios where computational resources are constrained, and global data validation is infeasible.

Future improvements could focus on refining the weighting signals, expanding validation diversity, or combining this approach with more selective filtering mechanisms to enhance discrimination between benign and adversarial updates. Ultimately, this work contributes toward the development of robust and deployable defenses for federated learning in adversarial environments.

References

[1] Zhang J, Wang J, Li H, et al. CHASe: Client Heterogeneity Aware Data Selection for Effective Federated Active Learning[J]. arXiv preprint arXiv:2504.17448, 2025.

[2] Nakano A, Chen S, Demachi K. Cross-Task Consistency Learning Framework for Multi-Task Learning[J]. Proceedings of Cityscapes & NYU Benchmark, 2021: alignment loss and cross-task consistency loss usage in multi-task setting.

[3] Li Q, He B, Song D. Model-Contrastive Federated Learning[J]. CVPR, 2021.

[4] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, Virginia Smith. Federated Learning: Challenges, Methods, and Future Directions[J]. Foundations and Trends in Machine Learning, 2021.

[5] H. Brendan McMahan et al. Communication-Efficient Learning of Deep Networks from Decentralized Data (FedAvg) [C]//2017.

[6] Burr Settles. Active Learning Literature Survey. University of Wisconsin-Madison.

[7] Jack Goetz, Kshitiz Malik, Duc Bui et al. Active Federated Learning. arXiv:1909.12641, 2019.

[8] Yu-Tong Cao, Ye Shi, Baosheng Yu et al. Knowledge-Aware Federated Active Learning with Non-IID Data. ICCV 2023.

[9] Goetz J, Malik K, Bui D, et al. Active Federated Learning[J]. arXiv preprint arXiv:1909.12641, 2019.

[10] Tang Y, Ren C, Tang X, et al. Efficient Heterogeneity-Aware Federated Active Data Selection (FALE)[P]. ICML 2025.

[11] "The Exploration–Exploitation Dilemma", PLOS ONE; overview of exploration-exploitation tradeoff in learning.

[12] Research Blog. Samsung Research. FedMargin: Federated Learning via Attentive Margin of Semantic Feature Representations[J]. 2023.

[13] Panagiotis, G. Ipeirotis, F. Provost, V. S. Sheng, and J. Wang, “Repeated labeling using multiple noisy labelers,” Data Min. Knowl. Discov., vol. 28, pp. 402–441, 2014.

[14] Wei X, et al. Optimizing Federated Learning by Entropy-Based Client Selection. arXiv preprint arXiv:2411.01240, 2024.

[15] Liu S, Deng W. Very deep convolutional neural network based image classification using small training sample size. IEEE ACPR, 2015.