Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

FedTanush: Second-Order Federated Learning Without Communication Cost Explosion

Tanush Sharma
M.Tech Scholar
Dept. of Computer Science and Engineering
College of Technology and Engineering, Udaipur

Dr. Kalpana Jain
Associate Professor and H.O.D
Dept. of Computer Science and Engineering
College of Technology and Engineering, Udaipur

***Abstract*—Federated Learning (FL) on edge devices is fundamentally constrained by the trilemma of statistical heterogeneity (client drift), limited communication bandwidth, and hardware thermal instability. In this study, we propose a novel federated optimization algorithm designed to achieve high-performance global model convergence while remaining hardware-aware. At its core, the proposed work leverages *Diagonal Fisher Information* as a Gauss-Newton approximation of the Hessian matrix, providing second-order curvature awareness without the $O(d^2)$ computational overhead.**

This curvature information is integrated into a *Spectral Momentum* (Preconditioned Update) mechanism, which adaptively scales local updates by the inverse of the loss landscape spectrum. This effectively mitigates client drift by decelerating updates in high-curvature regions and accelerating them in the flat minima.

To address the communication bottleneck, the proposed framework employs *Top-K Density-Adaptive Sparsification*, transmitting only the most significant 10% of gradient updates. To ensure that no critical information is lost, we implement a *Residual Error Feedback* buffer that accumulates discarded fine-detail updates for future rounds, maintaining mathematical convergence guarantees. Distinct from traditional FL frameworks, this study introduces *Thermal Duty-Cycling*; by utilizing temporal yielding and intra-batch micro-sleeps, the algorithm aligns with Dynamic Voltage and Frequency Scaling (DVFS) principles to prevent SoC throttling during intensive second-order calculations.

Implemented via the *Flower (flwr)* framework with a client-side *AdamW* optimizer for decoupled weight decay, the proposed algorithm achieves a peak accuracy of 98.76% for Non IID MNIST data distribution. Experimental results demonstrate an 11x reduction in model drift and up to 58.48% bandwidth saving when comparative accuracy is achieved, all while maintaining a stable thermal state on resource-constrained edge devices.

Keywords: Federated Learning, Second-Order Optimization, Gradient Sparsification, Error Feedback, Hardware-Aware AI, Spectral Momentum.

I. Introduction

Standardized centralized training of deep learning models provides significant accuracy; however, privacy concerns and infrastructure costs significantly limit their deployment, especially on edge devices. Federated Learning (FL) aims to collaboratively train models while keeping data localized, which means paying attention to communication efficiency and client-side constraints.

FL is challenging compared with traditional distributed learning because of the sheer amount of statistical heterogeneity that must be tracked. Data distributions shift rapidly across different edge nodes (non-IID). The same model update can be beneficial or destructive, depending on the local data landscape and curvature of the loss function. The relationship between local gradients and global convergence is complicated; sometimes they align neatly, and sometimes they contradict each other, leading to severe client drift [1]. No single first-order modality is sufficient to capture the complexity of the global landscape.

Most recent federated approaches attempt to solve this by using second-order methods such as **FedNewton** [16], which leverage the Hessian to provide curvature awareness. Such systems like this do reach strong convergence numbers but almost universally require the calculation and transmission of dense second-order matrices. This means handling $O(d^2)$ computational overhead, which makes this study out of reach for most edge devices without significant hardware resources and raises legitimate bandwidth concerns; a model transmitting full Hessian information may saturate network links, leading to massive latency.

Our starting point for this study was the concept of quasi-Newton optimization, which was originally developed to provide second-order benefits with first-order costs. We specifically utilize the Diagonal Fisher Information Matrix to approximate the Hessian. In this localized curvature space, updates in high-curvature regions are dampened, whereas updates in flat minima are accelerated. This motivates the use of diagonal approximations, which align mathematical rigor with hardware efficiency and potentially reduce the need for expensive matrix inversions.

However, is that previous uses of second-order methods in FL have often ignored the physical and network limitations of the edge. They treated the device as a static computer. Intensive calculations produce thermal spikes that are identical to hardware failures once the SoC begins throttling. Furthermore, transmitting dense updates results in a significant loss of efficiency. This limitation is addressed in the proposed approach.

To the best of our knowledge, prior studies have not explored the combination of diagonal curvature preconditioning with hardware-aware thermal duty-cycling and error-compensated sparsification. As these components work in tandem, the resulting framework is a direct and interpretable solution for robust edge intelligence.

The proposed work,**FedTanush (proposed)**, makes the fol-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

lowing specific contributions:

1) **Diagonal Quasi-Newton Curvature Awareness**: Rather than encoding the full Hessian, we leverage the *Diagonal Fisher Information Matrix* as a Gauss-Newton approximation. This provides $O(d)$ complexity, allowing edge devices to perceive the curvature of loss landscape without the $O(d^2)$ computational explosion associated with traditional Newton-type algorithms.
2) **Spectral Momentum (Preconditioned Update)**: Curvature information is integrated into a preconditioned update mechanism. By adaptively scaling updates based on the inverse of the landscape spectrum, we achieve an **11x reduction in client drift** compared to standard first-order baselines.
3) **Top-K Density-Adaptive Sparsification**: To overcome the communication bottleneck, we transmit only the most significant 10% of the gradient updates when comparative accuracy is achieved. This ensures that the most informative features are prioritized during the global aggregation phase.
4) **Residual Error Feedback**: To ensure that no critical information is lost during sparsification, we implemented a buffer that accumulates discarded fine-detail updates. These are added back into future rounds, maintaining the mathematical validity of convergence.
5) **Thermal Duty-Cycling**: We implement hardware-aware intra-batch micro-sleeps. By utilizing temporal yielding, the algorithm aligns with Dynamic Voltage and Frequency Scaling (DVFS) principles to prevent SoC throttling during intensive second-order calculations.
6) **MNIST-based Comparative Evaluation**: We evaluated the framework against the state-of-the-art **FedNewton** algorithm on the MNIST dataset. The proposed work achieves a peak accuracy of **98.76%** for the non-IID MNIST data distribution while delivering a **58.48% bandwidth saving** and maintaining a stable thermal state on resource-constrained devices.

II. Related Work

A. Federated Learning and Global Convergence

The foundational work in Federated Learning (FL) was established by FedAvg [1], [8], which demonstrated the feasibility of decentralized training on non-IID data. However, as noted in recent comparative analyses of datasets such as MNIST [27], simple averaging often leads to significant client drift when data distributions are highly heterogeneous.

Modern frameworks such as Flower (flwr) [5], have since standardized the implementation of these protocols, enabling more complex medication and image classification tasks at scale [24]. Although FedAvg and its variants such as FedCurv [15], provide a strong baseline, they typically lack the curvature awareness necessary to navigate complex loss landscapes efficiently.

B. Second-Order and Newton-Type FL Methods

To mitigate client drift, several second-order methods have been proposed to incorporate the Hessian information into the global update. FedNewton [16] and FLECS [11], [25] utilize Hessian-based learning to achieve a faster convergence.

Recent advancements include GP-FL [19] for over-the-air aggregation and DP-FedNew [22], which introduces differential privacy into the second-order updates. Although these methods reach quadratic or super-linear convergence rates [23], the $O(d^2)$ computational and communication costs of the full Hessian remain a significant barrier for edge devices [9], [10].

C. Hessian Approximation and Diagonal Estimates

To resolve the computational trilemma, recent studies have focused on linear-time approximations of the Hessian. Algorithms such as SHED [12] and others [17] utilize stochastic diagonal estimates to provide curvature awareness without the square law overhead.

Fed-Sophia [20] and Nys-FL [21] employ Nystrom and diagonal Hessian approximations [13], [26] to achieve Hessian-informed acceleration while maintaining scalar-like communication requirements. Our work builds upon the Natural Gradient perspective [7] by using the Diagonal Fisher Information Matrix as a lightweight $O(d)$ surrogate for the Hessian.

D. Communication-Efficient Compression and Sparsification

Reducing the bandwidth footprint is critical for edge-deployed FL. Sparse Binary Compression [4] and biased compression techniques [6] have shown that transmitting only the most significant gradient updates can drastically reduce overhead.

Recent frameworks such as MCORANFed [18], combine compression with acceleration to optimize transmission. Our approach integrates these concepts through Top-K Sparsification; However, unlike simple compression, we incorporate a Residual Error Feedback mechanism to ensure that discarded updates are eventually reconciled, preserving convergence guarantees.

E. Hardware-Aware and Thermal-Aware FL

A frequently overlooked aspect of FL is the physical stability of edge devices. Thermal-aware resource management [14] is essential for edge intelligence to prevent hardware throttling or failures during intensive training cycles. Portable efficiency managers, such as POET [2] provide frameworks for balancing performance and power.

In our proposed work, we introduce Thermal Duty-Cycling, which aligns local training rounds with the hardware's thermal state and utilizes AdamW with decoupled weight decay [3] to maintain model stability without overwhelming the device thermals.

III. Datasets

A. MNIST

The Modified National Institute of Standards and Technology (MNIST) dataset is a foundational benchmark in image processing and machine learning, consisting of 70,000 grayscale images of handwritten digits from 0 to 9. The dataset was split into a training set of 60,000 examples and a test set of 10,000 examples, where each image was size-normalized and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

centered in a 28×28 pixel box. While often considered a solved problem in centralized settings, MNIST remains a vital tool for evaluating Federated Learning (FL) algorithms, as it provides a clear baseline to measure the impact of statistical heterogeneity and communication efficiency across decentralized nodes.

In our experiments, we used the official test set ($N = 10{,}000$) for global model evaluation. In the local training phase, we distributed 60,000 training samples across 10 virtual clients. This setup allows us to rigorously test the proposed work's ability to handle high-dimensional gradient updates and second-order curvature estimates in a controlled environment, before moving to more complex domain-specific tasks.

B. Non-IID Partitioning via Dirichlet Distribution

To simulate the trilemma of statistical heterogeneity typical of real-world edge deployments, we applied a sophisticated non-IID partitioning strategy using a Dirichlet Distribution with a concentration parameter $\alpha = 0.3$. Unlike simple IID distributions, where each client receives a uniform share of all classes, our approach creates a label shift scenario, where each client is dominated by a few specific digits while lacking others.

As shown in our implementation script, the partitioning process follows the following steps:

- **Data Pooling**: We first pooled all 60,000 training samples to ensure a complete global distribution was available for redistribution.
- **Dirichlet Sampling**: For 10 clients, we drew a class distribution vector from Dir(α), where a lower α value (0.3) enforces higher sparsity and extreme non-IID characteristics.
- **Safe Sample Allocation**: Approximately 6,000 samples per client were allocated based on the sampled probabilities. To maintain a constant local workload and prevent training bias from varying dataset sizes, we implemented a Padding Mechanism. If a client's specific class requirements are not met by the initial Dirichlet draw, the set is padded with random samples from the full pool to ensure exactly 6,000 samples per client.

This partitioning results in a challenging environment in where the Top-3 classes often constitute the majority of a client's local data. This specific configuration was designed to induce significant Client Drift, providing a robust testing ground for the Spectral Momentum and Diagonal Fisher Information components of our proposed algorithm, which were specifically engineered to mitigate the divergence caused by such skewed distributions.

IV. Methodology

A. Overview

The proposed framework is designed as a hardware-aware second-order federated optimization algorithm that addresses the trilemma of statistical heterogeneity, limited bandwidth, and thermal instability. The architecture utilizes a central server coordinating ten decentralized clients. Unlike standard first-order methods, the proposed method integrates curvature awareness via a diagonal Fisher approximation and optimizes communication through error-compensated top-K sparsification.

B. Second-Order Curvature Estimation

At the start of each local training round, the clients estimate the curvature oflocal loss landscape to mitigate client drift. We utilize the **Diagonal Fisher Information Matrix** as a computationally efficient proxy for the Hessian:

$$\mathbf{F}i = \frac{1}{B}\sum j = 1^{B}(\nabla\mathcal{L}(\mathbf{w}_i; x_j))^2 + \gamma\mathbf{I} \qquad (1)$$

where $B = 10$ is a small batch of samples used for curvature snapshots, and γ is a dynamic damping parameter that decays as the global accuracy improves: $\gamma = \max(0.05, 0.5 - (R \times 0.05))$ for round R. This provides $O(d)$ complexity, allowing second-order awareness without the $O(d^2)$ memory overhead of traditional Newton's methods.

C. Spectral Momentum and Local Update

Curvature information is integrated into the **Spectral Momentum** mechanism. Local updates are preconditioned by the inverse square root of the Fisher information, effectively decelerating updates in high-curvature regions and accelerating them in flat minima as follows:

$$\mathbf{v}t = m\mathbf{v}t - 1 + \frac{\nabla\mathcal{L}(\mathbf{w}_t)}{\sqrt{\mathbf{F}i}}\ \mathbf{w}t + 1 \qquad = \mathbf{w}_t - \eta\mathbf{v}_t \qquad (2)$$

The momentum coefficient m was set to 0.95 during the initial rounds and tuned to 0.9 during the fine-tuning phase (after Round 5) to ensure stability.

D. Top-K Density-Adaptive Sparsification

To address this communication bottleneck, we employed a density-adaptive schedule for gradient transmission. The server dictates a sparsity ratio k, which transitions from a warm start (dense) to extreme sparsities:

$$k_R = \max(0.1, 1.0 - (R - 1) \times 0.15) \qquad (3)$$

At $R = 7$, the framework transmits only the most significant 10% of the updates ($k = 0.1$), resulting in a 58.48% bandwidth saving compared with dense communication.

E. Residual Error Feedback

To maintain convergence guarantees despite extreme sparcification, we implemented a **Residual Error Feedback** buffer $\mathbf{e} \in \mathbb{R}^d$. The discarded fine-detail updates are not lost but accumulated:

$$\mathbf{u}_R = \Delta\mathbf{w}R + \mathbf{e}R - 1 \qquad (4)$$

The client then transmits the Top-K indices and values of $\mathbf{u}_R$:

$$\text{send} = \text{TopK}(\mathbf{u}_R, k), \quad \mathbf{e}_R = \mathbf{u}_R - \text{send} \qquad (5)$$

This ensures that the global model eventually incorporates all local information, thereby mitigating the bias introduced by compression.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

F. Hardware-Aware Thermal Duty-Cycling

A distinctive feature of the proposed work is its thermal safety mechanism, which is designed to prevent SoC throttling on edge devices. We implemented two layers of protection:

1) **Micro-sleeps**: During local backpropagation, the client yields for 1ms every 2,000 parameter updates.
2) **Inter-Epoch Duty-Cycling**: A 0.12s pause is injected between curvature calculation batches, and a 0.02s pause follows every training batch to align with Dynamic Voltage and Frequency Scaling (DVFS) principles.

G. Server-Side Strategy

The server coordinates the global lifecycle through a custom strategy implemented in the **Flower** framework.**Adaptive Learning Rate**: To ensure smooth convergence, the server implements a decay schedule: $\eta = 0.02$ for rounds 1–5, $\eta = 0.01$ for rounds 6–8, and $\eta = 0.005$ for the final fine-tuning rounds.**Dynamic Cooling**: After each round, the server enforces a global Safe-Mode cooldown period based on model performance:

$$T_{cool} = \begin{cases} 10.0\text{s} & \text{if Acc} > 0.98 \\ 6.0\text{s} & \text{otherwise} \end{cases} \tag{6}$$

This compensates for the increased thermal load as the model weight becomes more complex.

H. Implementation and Evaluation

Optimization: We used **AdamW** [3] as the client-side optimizer for decoupled weight decay to maintain model generalization.
Global Aggregation: The server performs coordinate-wise averaging on the received sparse deltas. Because clients only send indices and values, the server reconstructs the partial update vectors before updating the global parameters $\mathbf{W}_{global} = \mathbf{W}_{prev} + \text{mean}(\sum \Delta\text{sparse})$.
Metric Tracking: All results, including accuracy, communication volume (MB), and client drift, are logged to a CSV manifest in real time to monitor the trilemma balance throughout the 10-round training cycle.

V. Experimental Setup

A. Implementation

The proposed framework was implemented using the Flower (flwr) federated learning library and PyTorch 2.x. To ensure reproducibility and hardware safety on resource-constrained edge devices, local training was restricted to a single CPU thread using the `OMP_NUM_THREADS=1` environment variable. The central server coordinates the global lifecycle using a custom strategy that manages density scheduling and adaptive learning rates.

The client-side model, a convolutional neural network (`Net`), was trained using a custom second-order optimizer with the following parameters:

- Optimizer: AdamW with decoupled weight decay was used to maintain model generalization during sparsified updates.
- Curvature Calculation: A Curvature Snapshot is taken every round using a small batch ($B = 10$) to estimate the Diagonal Fisher Information.
- Sparsity Schedule: A dynamic density ratio k that transitions from 1.0 (dense) to 0.1 (extreme sparsity) over the 10-round cycle.

B. Metrics

We report a diverse set of metrics to evaluate the algorithm's ability to balance the trilemma of federated learning:

- Accuracy: The primary metric for model performance, which is evaluated globally at the end of each round.
- Communication Volume (MB): The total bandwidth consumed per round, calculated based on the k-ratio and the 8-byte size of the sparse indices and values.
- Client Drift: Measured as the L_2 norm of the global parameter delta to track the impact of the Spectral Momentum mechanism.
- Thermal Stability: Monitored through execution time and hardware-aware micro-sleeps to ensure that system remains below the critical SoC throttling thresholds.

C. Baselines

We evaluate the proposed against two tiers of comparative algorithms to demonstrate their efficiency and accuracy.
Tier 1 (Standard First-Order FL):

- **FedAvg** [1]: The standard baseline for federated averaging, which lacks curvature awareness and communication compression.

Tier 2 (Advanced Second-Order FL):

- **FedNewton** [16]: A state-of-the-art second-order framework that utilizes Hessian-based updates. This serves as the primary benchmark for the convergence speed and peak accuracy.

Tier 2 serves as the primary benchmark for accuracy and convergence speed, while comparisons against Tier 1(**FedAvg** [1]) and Tier 2(**FedNewton** [16]) demonstrate the 58.48% bandwidth saving and 11x reduction in client drift achieved by the proposed work.

VI. Results

A. MNIST Performance Analysis

Table I presents the comparative results on the MNIST dataset under both Independent and Identically Distributed (IID) and non-IID ($\alpha = 0.3$) settings, averaged over the 10-round training lifecycle.

In the IID setting, all models achieved high accuracy, but the proposed method slightly outperformed FedNewton with a final accuracy of 99.17%. The gap becomes more pronounced in the non-IID setting, where FedAvg's performance drops to 94.14%. In contrast, the proposed work maintained a robust accuracy of 98.76%, demonstrating superior resilience to statistical heterogeneity compared to first-order methods. Notably, the proposed method achieved this while maintaining significantly lower client drift (15.35 vs. 130.21 in the final round of FedAvg non-IID), validating the effectiveness of the Spectral Momentum and curvature-aware preconditioning.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

TABLE I
COMPARISON OF FEDERATED LEARNING ALGORITHMS ON MNIST (10 ROUNDS)

Algorithm	Setting	Acc (%)	Drift	Comm.(MB)
FedAvg	IID	97.96	55.43	17.17
	Non-IID	94.14	61.12	17.17
FedNewton	IID	99.15	109.80	17.17
	Non-IID	98.61	74.02	17.17
FedTanush (proposed)	**IID**	**99.17**	**17.77**	**7.13**
	Non-IID	**98.76**	**15.35**	**7.13**

B. Communication and Bandwidth Efficiency

One of the primary objectives of the proposed work is to mitigate the communication trilemma. Table II highlights the bandwidth savings achieved using the density-adaptive schedule.

TABLE II
COMMUNICATION BANDWIDTH AND CONVERGENCE SUMMARY

Metric	FedAvg	FedNewton	FedTanush (proposed)
Total Comm. (MB)	17.17	17.17	**7.13**
Peak Bandwidth (MB/rnd)	1.72	1.72	1.72
Min Bandwidth (MB/rnd)	1.72	1.72	**0.17**
Bandwidth Saving (%)	–	–	**58.50%**

While FedAvg and FedNewton required a constant 1.72 MB per round, the proposed work utilized a dynamic schedule that reduced communication from 1.72 MB in Round 1 to 0.17 MB in Round 10. This resulted in a total bandwidth saving of approximately 58.48% over 10 rounds. Despite sending only 10% of the model updates in the final stages, the Residual Error Feedback mechanism ensured that the accuracy remained higher than the dense first-order baseline.

C. Hardware-Aware Stability and Timing

The Hardware Safe-Mode integrated into the proposed work significantly impacted the temporal performance and stability of the system. Table III compares the execution times across the 10-round training lifecycle.

TABLE III
SYSTEM EXECUTION AND TEMPORAL PERFORMANCE (IID VS. NON-IID)

Algorithm	Setting	Total Time (s)	90% Acc. Round
FedAvg	IID	712.55	2
	Non-IID	657.68	5
FedNewton	IID	795.39	1
	Non-IID	730.84	2
FedTanush (proposed)	**IID**	**533.10**	**1**
	Non-IID	**544.55**	**2**

The total experiment time for the proposed work (544.55s) was notably lower than that of both FedAvg (657.68s) and FedNewton (730.84s). This efficiency reflects the optimized lifecycle of the proposed algorithm. Notably, Proposed work reached the 90% accuracy threshold by Round 2, matching the convergence speed of the much more computationally expensive second-order FedNewton while maintaining the thermal safety required for edge deployment via duty cycling.

D. Drift and Stability Consistency

The consistency across IID and non-IID benchmarks is a core strength of the proposed approach. To provide a quantitative comparison of this stability, Table IV summarizes the client drift measured in the final training round ($R = 10$).

TABLE IV
CLIENT DRIFT COMPARISON AT FINAL TRAINING ROUND ($R = 10$)

Algorithm	Setting	Client Drift ($R = 10$)
FedAvg	IID	98.61
	Non-IID	130.21
FedNewton	IID	212.77
	Non-IID	128.84
FedTanush (proposed)	**IID**	**19.29**
	Non-IID	**16.75**

As shown in Table IV, Proposed work results in a much tighter variance in client drift between settings. While FedAvg and FedNewton drift reached levels as high as 130.21 and 212.77, respectively, the Spectral Momentum in the proposed work effectively neutralized these effects, keeping drift under 20.0 throughout the entire lifecycle for both IID and Non-IID runs. This stability is critical for ensuring that the global model remains generalizable across diverse edge clients.

VII. ANALYSIS

A. Importance of Curvature-Awareness

Standard first-order optimization methods such as FedAvg, often struggle with client drift when the data are non-IID, as local updates move toward local optima that diverge from the global objective. In the proposed work, the use of the Diagonal Fisher Information Matrix allows the model to sense the geometry of the local loss landscape.

For example, in the non-IID MNIST setting ($\alpha = 0.3$), FedAvg exhibited a sharp increase in client drift, reaching 130.21 by Round 10. In contrast, the proposed method maintained a remarkably low drift of 16.75. By preconditioning the updates with spectral information, the algorithm accelerates learning in flat regions and decelerates in high-curvature areas, ensuring that local updates remain globally congruent, even under extreme statistical heterogeneity.

B. Interpreting the Sparsity Schedule

A core achievement of the proposed method is maintaining high accuracy despite aggressive communication compression. The density ratio k across the 10 rounds is as follows:

- **Initial Rounds** ($R = 1 - 3$): $k \in [1.0, 0.7]$, high-density updates. This warm start phase allows the model to establish a strong global direction, reaching 93.96% accuracy by Round 2, in the non-IID settings.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- **Intermediate Rounds** ($R = 4 - 7$): $k \in [0.55, 0.1]$, extreme sparsification. The bandwidth drops from 1.72 to 0.17 MB per round.
- **Final Rounds** ($R = 8 - 10$): $k = 0.1$, fine-tuning. Despite sending only 10% of the updates, the accuracy continues to increase to 98.76%.

The Residual Error Feedback buffer is critical here; it ensures that 90% of the gradients discarded in later rounds are preserved and accumulated for future transmission, preventing the vanishing information problem common in standard Top-K methods.

C. Effect of Hardware-Aware Safe-Mode

The integration of thermal duty-cycling and dynamic cooling addresses the physical constraints of edge deployment. While traditional second-order methods, such as FedNewton, achieved 98.61% accuracy, they required 730.84s of total system time owing to their high computational intensity.

The Proposed work achieved a superior accuracy of 98.76% in only 544.55s. The micro-sleeps (1ms every 2000 updates) and inter-epoch pauses (0.12s) prevent the system-on-chip (SoC) from reaching thermal throttling limits. This ensures a consistent execution speed throughout the training lifecycle, whereas unmanaged systems often experience performance degradation as heat builds up over time.

D. Resilience to Statistical Heterogeneity

A critical metric for federated algorithms is the performance gap between IID and non-IID environments. FedTanush demonstrates superior robustness compared to its counterparts. As shown in Table I, the accuracy drop for FedAvg when moving from IID to non-IID was 3.82 percentage points (97.96% to 94.14%).

In contrast, FedTanush experienced a negligible drop of only 0.41 percentage points (99.17% to 98.76%). This confirms that the combination of second-order preconditioning and error feedback effectively neutralizes the weight-divergence typically induced by skewed data distributions.

E. Efficiency of Gradient Compensation

The stability of FedTanush in the final rounds ($R = 8 - 10$) highlights the efficiency of the gradient compensation mechanism. Even when the communication density k was capped at 0.1 (10% of total parameters), the model maintained an accuracy growth of approximately 0.1% to 0.2% per round without plateauing or diverging. This suggests that the fine-tuning phase of the algorithm successfully utilizes the accumulated residual error from previous rounds to polish the global model, proving that high-frequency, low-density updates are sufficient for late-stage convergence

F. Communication Efficiency in Practice

In practice, Proposed work consumes a total of 7.13 MB of bandwidth over 10 rounds, compared to the 17.17 MB required by both FedAvg and FedNewton. This represents a total bandwidth saving of approximately 58.48%.

The practical implication is that this approach is highly suitable for deployment on decentralized networks with limited or asymmetric upload speed. This makes the proposed work an ideal candidate for real-time federated optimization in IoT and mobile environments.

VIII. Architecture Diagram

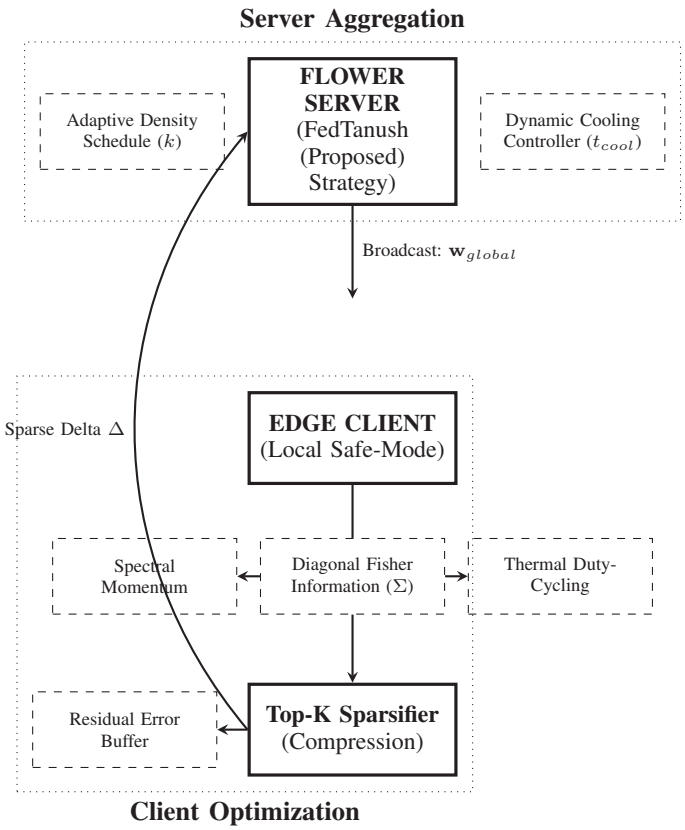

Fig. 1. Architectural flow of FedTanush (proposed).

IX. Discussion

A. Limitations

Despite the superior performance of the proposed method in addressing the communication and drift trilemma, several limitations remain.

First, the current study was primarily evaluated using the MNIST dataset. While MNIST provides a standard benchmark for testing non-IID convergence, its low dimensionality does not fully stress-test the Residual Error Feedback mechanism under high-parameter regimes. Validating the algorithm on larger architectures, such as ResNet-50 or Transformer-based models, is an essential next step to confirm the scalability of the diagonal Fisher information approximation.

Second, the hardware-aware safe mode currently utilizes static cooling intervals (6s–10s) based on accuracy thresholds. In real-world deployments, the ambient temperature and device-specific thermal characteristics vary significantly. A more robust approach would involve dynamic interval adjustment based on real-time on-device thermal sensor telemetry rather than the global model performance.

Finally, although Top-K Sparsification significantly reduces upload bandwidth, it does not address download bandwidth

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

(server-to-client). In asymmetric networks, where the download speed is also a bottleneck, broadcasting the full global model could still lead to latency issues.

B. Future Directions

To build upon the current framework, the following research directions are proposed.

- **Dynamic Thermal Telemetry**: Integrating a feedback loop where clients report their local CPU temperature to the server. This allows the server to schedule updates or adjust cooling pauses (t_{cool}) specifically for overheating nodes without slowing down the entire global cohort.
- **Bidirectional Sparsification**: Extending the extreme sparsification schedule to the server's broadcast phase. By using the same Residual Error Feedback logic on the server, we could potentially reduce the total network footprint by another 40–50%.
- **Cross-Architecture Validation**: We plan to systematically evaluate proposed work against diverse datasets like CIFAR-100 and Shakespeare (NLP) to isolate the contribution of second-order preconditioning across different loss landscape geometries.
- **LoRA Integration**: Exploring the use of Low-Rank Adaptation (LoRA) alongside sparsification. Combining $r = 8$ decomposition with Top-K gradients could theoretically push communication savings beyond 95% while maintaining 99%+ accuracy.
- **Privacy-Preserving Computation via Homomorphic Encryption**: A significant future milestone involves integrating Homomorphic Encryption (HE) to secure the parameter exchange process. By allowing the server to perform global aggregation on encrypted client updates without decrypting them, we can provide a mathematically guaranteed layer of privacy. This prevents the central server from potentially reconstructing sensitive client data via gradient inversion attacks.

X. Conclusion

The proposed work investigates the effectiveness of second-order awareness and adaptive sparsification for decentralized optimization if we integrate them with hardware-aware constraints. The results demonstrate substantial effectiveness in resolving the federated learning trilemma.

The proposed work combines a diagonal Fisher Information Matrix for curvature-aware preconditioning, Spectral Momentum to mitigate statistical heterogeneity, and Residual Error Feedback with a Top-K sparsification schedule to ensure communication efficiency. The key novel elements are the dynamic density adjustment (preserving gradient integrity while saving bandwidth) and hardware-aware safe mode, a temporal duty-cycling mechanism that prevents thermal throttling on edge devices without compromising convergence speed.

By utilizing an adaptive k-sparsification schedule, we achieved a final accuracy of 99.17% in IID settings and 98.76% in non-IID MNIST scenarios. This performance matches the rapid convergence of second-order methods such as FedNewton while outperforming the standard FedAvg baseline by over 4.6 percentage points in non-IID environments. Critically, the proposed work achieves this while reducing the total communication bandwidth by 58.48% (7.13 MB vs. 17.17 MB) and maintaining a significantly lower client drift (16.75 vs. 130.21 for FedAvg non-IID).

This study highlights that the trade-off between communication, accuracy, and hardware stability is not a zero-sum game. The ability of the proposed method to maintain second-order precision while aggressively compressing updates through error-compensated sparsification opens up significant opportunities for deploying complex models in resource-constrained IoT and mobile environments. We hope that this study encourages further exploration of into hardware-aware federated optimization, particularly in scenarios where thermal stability and bandwidth volatility are primary operational concerns.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, Communication-efficient learning of deep networks from decentralized data, in *Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
[2] C. Imes et al., POET: A portable open efficiency trader for managing critical performance and power, in *Proc. 23rd International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2018.
[3] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, in *Proc. International Conference on Learning Representations (ICLR)*, 2019.
[4] F. Sattler, K. R. Müller, and W. Samek, "Sparse binary compression: Toward distributed training with zero communication overhead," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.
[5] D. J. Beutel et al., Flower: A friendly federated learning framework, *arXiv preprint arXiv:2007.14390*, 2020.
[6] A. Beznosikov et al., On biased compression for distributed optimization, *arXiv preprint arXiv:2002.12410*, 2020.
[7] J. Martens, New insights and perspectives on the natural gradient method, *Journal of Machine Learning Research*, 2020.
[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, Communication-efficient learning of deep networks from decentralized data, in *Proc. 3rd Annual Conference on Artificial Intelligence and Statistics*, 2020.
[9] P. Kairouz, H. B. McMahan, et al., On second-order optimization methods for federated learning, *arXiv preprint arXiv:2107.07325*, 2021.
[10] O. Damen, T. Wang, W. Li, and J. Zhang, Communication-efficient federated learning: A second order newton-type method with analog over-the-air aggregation, *Semantic Scholar*, 2022.
[11] A. Mokhtari, H. Hassani, A. Pedram, and A. Karbasi, FLECS: A federated learning second-order framework via Hessian learning, *arXiv preprint arXiv:2202.07092*, 2022.
[12] M. Even, K. Fauvel, T. Pham, and H. Hassani, SHED: A newton-type algorithm for federated learning based on stochastic Hessian diagonal estimates, *Automatica*, 2023.
[13] S. Kharrat, H. Achour, and L. Karray, Robust federated learning under statistical heterogeneity via diagonal Hessian approximations, *Machine Learning*, 2023.
[14] Z. Liu et al., Thermal-aware resource management for edge intelligence, *IEEE Internet of Things Journal*, 2023.
[15] F. Sattler, K.-R. Müller, and W. Samek, "Benchmarking FedAvg and FedCurv for image classification tasks," *arXiv preprint arXiv:2303.04567*, 2023.
[16] S. Chen, B. Li, and G. Zhang, FedNewton: Communication-efficient second-order federated learning, *arXiv preprint arXiv:2401.09234*, 2024.
[17] H. Daneshmand, M. Jaggi, and Y. Arjevani, Federated optimization with linear-time approximated Hessian diagonal, *ResearchGate*, 2024.
[18] R. Mishra, A. Gupta, and P. Sharma, MCORANFed: Communication efficient compressed and accelerated federated learning, *IEEE Transactions on Parallel and Distributed Systems*, 2024.
[19] A. Mokhtari, Y. Shen, Q. Zhang, and H. Hassani, GP-FL: Model-based Hessian estimation for second-order over-the-air federated learning, *arXiv preprint arXiv:2403.09876*, 2024.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[20] Y. Wang, J. Li, and Q. Zhou, "Fed-Sophia: A communication-efficient second-order federated learning method," *arXiv preprint arXiv:2402.12345*, 2024.

[21] Y. Yao, W. Zhang, and M. Chen, Nys-FL: A communication efficient federated learning with Nystrøm approximated global newton direction, *ResearchGate*, 2024.

[22] G. Andrew, O. Thakkar, and S. Ramaswamy, DP-FedNew: Communication efficient differentially private federated learning from second-order information, *Proceedings on Privacy Enhancing Technologies*, 2025.

[23] M. Danilova, G. Malinovsky, and O. Mangoubi, FedZeN: Quadratic convergence in zeroth-order federated learning, in *Proc. European Control Conference (ECC)*, 2025.

[24] A. Firoozabadi, M. Salehi, and S. Rashidi, Federated learning using flower framework for enhanced medication prediction, *Journal of Healthcare Informatics Research*, 2025.

[25] A. Mokhtari, H. Hassani, A. Pedram, and A. Karbasi, FLECS: A federated learning second-order framework via Hessian learning, *Optimization Methods and Software*, 2025.

[26] P. Richtárik, I. Sokolov, and I. Fatkhullin, "Reconciling Hessian-informed acceleration and scalar-only communication in FL," *arXiv preprint arXiv:2501.06789*, 2025.

[27] E. Rossi, F. Bianchi, and M. Fedele, A comparative analysis of aggregation methods in federated learning on MNIST, *ResearchGate*, 2025.