Effects of Send-on-Delta Sampling on Federated Human Activity Recognition

Pranav Chapa
Independent Researcher, Prosper, Texas, USA
chapapranav2@gmail.com

***Abstract*—Federated human activity recognition usually treats sensor acquisition as fixed: every client first records a periodic stream, after which federated learning begins. Edge devices may instead sup-press locally redundant measurements through signal-dependent sam-pling, and different clients may apply different thresholds because of bandwidth or sensing constraints. This paper formally defines acquisi-tion-policy heterogeneity for horizontal federated classification and derives application requirements for cost isolation, recognition utility, client-group reliability, and policy-shift transparency. Send-on-Delta sampling is evaluated against periodic and random retention at exactly matched within-window packet counts using a leakage-safe protocol, twelve paired optimization seeds, and a fixed round-30 endpoint on WISDM, PAMAP2, and MotionSense. At 50.5% retention on WISDM, Send-on-Delta improves the test-sample-weighted mean client Macro-F1 over matched periodic sampling by 0.0120 and over matched random retention by 0.0368; both paired effects remain significant after Holm correction for the two WISDM comparisons. PAMAP2 and MotionSense provide no evidence of a packet-count-matched advantage at 57.4% and 46.1% retention, respectively; their SOD Macro-F1 means differ from full periodic by -0.0013 and -0.0001, but no equivalence margin was predeclared. An eight-seed policy-shift audit exposes a WISDM failure boundary under aggressive post-training threshold increases; under the PAMAP2 severe policy, the mean changes by +0.0001 and its 95% confidence interval crosses zero. In sequence-state-ful sensitivities, the SOD-random intervals are positive on WISDM and MotionSense, while every SOD-periodic interval crosses zero. A prede-clared convolutional neural network with doubled channel widths and about 3.8 times as many trainable parameters preserves the WISDM SOD-periodic direction and yields only a small PAMAP2 effect. The reusable finding is conditional: retained packet count alone can be insuf-ficient to characterize sparse federated sensing, but the value of signal-dependent timestamps depends on the activity stream, model, and oper-ating policy.**

***Index Terms*—Federated learning, human activity recognition, Internet of Things, Send-on-Delta sampling, event-based sensing, edge intelligence.**

I. Introduction

Federated learning (FL) trains a shared model from decentralized client data without collecting raw examples at a central server [1]. This is attractive for wearable and mobile human activity recognition (HAR), where inertial streams are continuous and can reveal sensitive behavior. Prior federated HAR work addresses user-level signal heterogeneity and personalization after sensor windows have been collected [2]. Most evaluations nevertheless begin after acquisition: every client is assumed to possess a regularly sampled signal, and heterogeneity is introduced through subjects, labels, devices, or periodic sampling rates. This omits a pretraining systems decision: which physical measurements reach the edge classifier.

Send-on-Delta (SOD) reports a sample when its deviation from the last report reaches a threshold [3], [4]. It suppresses locally redundant measurements without waiting for a learned classifier and yields irregular, signal-dependent streams. In a federation, distinct client thresholds produce different retained timestamps and therefore different local learning distributions. This acquisition-policy heterogeneity is not fixed-rate downsampling or random missingness: silence means the signal remained inside a client-specific threshold band.

A. Application Setting and Requirements

The target setting is non-safety-critical wearable or phone monitoring of everyday activities. A low-power inertial unit supplies measurements to a colocated phone or gateway; that processor reconstructs fixed-duration windows, performs recognition and local training, and exchanges shared model state with a federated coordinator. SOD acts on the sensor-to-processor path, whereas FL acts on the processor-to-server path. Suppressing a model update leaves local examples intact; suppressing a measurement changes every later local update computed from that window.

This application analysis yields five requirements that are stated before the method and assessed quantitatively in Section VI. They are evaluation requirements, not claims of certified battery life or a universal minimum accuracy, because none of the datasets defines a product-level acceptance threshold.

1) R1, cost isolation: policies compared as sensing alternatives must retain exactly the same number of three-axis packets in every client window.
2) R2, recognition utility: Macro-F1, accuracy, paired uncertainty, and multiplicity-controlled significance must be reported at the fixed sensing budget; superiority is claimed only when the paired evidence supports it.
3) R3, client reliability: federation-level utility must be accompanied by threshold-group performance so that an average cannot hide a poorly served sensing-budget group.
4) R4, policy-shift transparency: post-training threshold changes must be evaluated as acquisition-distribution shifts, and failure boundaries must be reported rather than described as robustness.
5) R5, edge compatibility: all acquisition modes must use the same compact model and causal reconstruction, raw signals must remain local to the client, and retention must not be relabeled as measured energy savings.

B. Research Questions and Contributions

The main experimental difficulty is cost matching. If SOD retains more measurements than a baseline, higher accuracy may simply reflect a larger budget. Conversely, comparison only with the complete periodic stream cannot isolate evidence preserved per retained packet. This study asks whether SOD changes federated HAR relative to periodic and random timestamps at an identical within-window packet count, whether the effect repeats across three public datasets, and what happens when deployed thresholds differ from training.

The contributions are stated as two formal results and three transferable empirical findings:

1) C1: Equations (1)-(3) define acquisition-policy heterogeneity as client-specific acquisition and reconstruction operators that induce the datasets optimized by FL, separating sensor events from event-triggered model communication.
2) C2: Algorithm 1 gives an exact retained-count protocol in which SOD, periodic, and random streams share packet count, causal reconstruction, classifier, partitions, and optimization; zero count mismatches were found across 45,725 prepared windows.
3) F1: On WISDM, SOD produces significantly higher Macro-F1 than both packet-count-matched controls at 50.5% retention, showing that retained packet count alone need not characterize sparse federated sensing.
4) F2: PAMAP2 and MotionSense provide no evidence of a packet-count-matched SOD advantage at 57.4% and 46.1% retention, establishing an external-validity boundary rather than universal superiority.
5) F3: Across eight policy-shift seeds, severe post-training sparsity lowers WISDM Macro-F1 by .0669 and its weakest-group mean to .2227. The PAMAP2 severe-policy mean changes by +.0001 with a confidence interval that crosses zero, and untargeted compute-matched threshold augmentation harms WISDM.

This claim structure determines the remainder of the paper. Section III establishes C1 and C2, Section IV defines the controls needed to test them, Section V reports F1-F3, and Section VI evaluates those findings against R1-R5. Thus, the formal contributions, empirical findings, and application requirements are declared before their evidence is presented.

II. Related work

This section determines whether the study's claimed contribution is already covered by signal-dependent sensing or federated temporal-heterogeneity research. It therefore separates sensor-level event generation from event-triggered model communication and identifies the missing controlled comparison that motivates C1 and C2.

A. Signal-Dependent Sensing

SOD reports a new value when the deviation from the last transmitted value reaches a threshold [3]. Prior work analyzes its report traffic and tracking error, including dynamic threshold selection under a transmission budget [4]. Wearable sensing research has also reduced transmissions using local decision mechanisms and evaluated activity-recognition accuracy alongside network lifetime [5]. Adaptive-resolution HAR separately selects periodic sampling rates according to sample difficulty and computational budget [6]. These studies establish sensing, computation, and reconstruction trade-offs, but they do not examine collaborative model training across clients whose local streams were produced by different persistent thresholds.

Reconstruction quality and classification utility are related but not interchangeable. A sampler can minimize mean signal error while discarding brief transitions that distinguish two activities, or retain large changes that are irrelevant to a classifier. Centralized signal-to-event work has therefore selected SOD and related encoding parameters using downstream classification performance and event count [7]. The present protocol instead tests persistent client-specific thresholds in horizontal FL, does not tune them from test reconstruction error, and uses matched periodic and random controls rather than comparing only with a dense signal.

B. Temporal Heterogeneity in Federated Learning

Federated averaging combines locally optimized model states and is a standard baseline for decentralized data [1]. For feature shift, FedBN keeps batch-normalization state local rather than averaging it across clients [8]. This is relevant here because different acquisition thresholds alter local signal distributions even when the physical sensor type is shared.

Temporal heterogeneity has been studied through periodic sampling frequency. Multi-frequency FL-HAR supports head-worn devices operating at different fixed frequencies [9]. Recent work on federated spiking networks similarly adapts models across heterogeneous temporal resolutions [10]. These studies vary fixed sampling frequency or temporal binning rather than treating a persistent, client-specific SOD threshold as the acquisition policy. Earlier federated neuromorphic learning used dynamic vision sensor (DVS) event data with Send-on-Delta-style pixel encoding [11], but treated the encoded stream as a common input modality rather than a client-specific acquisition policy or experimental variable. In contrast, our SOD event times depend on each client's threshold, the signal trajectory, and the previous transmitted state, so two windows with the same retained count may preserve different timestamps.

The phrase event-triggered FL is also used when a device transmits model parameters only after a model-change criterion is met. Zehtabi et al. use heterogeneous communication thresholds for decentralized model exchange [12]. That mechanism changes when learning updates cross the network. Our threshold acts one layer earlier: it changes which sensor measurements constitute each client's local training and evaluation data. To the best of our knowledge, prior work has not systematically isolated this client-specific signal-dependent acquisition effect in horizontal federated classification using exact packet-count-matched baselines.

In summary, prior work establishes SOD, adaptive sensing, federated HAR, and temporal-resolution heterogeneity separately, but it does not isolate client-specific signal-dependent acquisition in horizontal FL at exactly matched packet counts. This gap motivates C1's formulation of acquisition-policy heterogeneity, C2's retained-count protocol, and the policy-shift analysis used to bound their practical interpretation.

III. Problem formulation

This section provides the two formal contributions claimed in Section I-B. It first defines how a client acquisition policy induces the local dataset optimized by FL, then specifies the SOD operator and the exact retained-count protocol required to isolate timestamp selection from sensing quantity.

A. Acquisition-Policy Heterogeneity

Intuitively, two clients can observe the same activity through the same sensor type yet train on different timestamps because their acquisition thresholds differ. The heterogeneity is introduced before model training: the acquisition policy determines the examples on which the local loss is evaluated.

Let X_{iw} be a regular $T\times3$ window w at client i, with label y_{iw}. A client policy $\pi_i=(A_i,R_i,\delta_i)$ contains an acquisition operator, reconstruction operator, and threshold. For policy set $\Pi=\{\pi_i\}_{i=1}^{N}$, the induced local dataset is

$$\widetilde{D}_i^{\pi_i} = \left\{ (\mathcal{R}_i(\mathcal{A}_i(X_{iw};\delta_i)), y_{iw}) \right\}_{w=1}^{n_i}. \quad (1)$$

Equation (1) is the referable formulation of acquisition-policy heterogeneity: $|\Pi|>1$, so at least two clients use different policies, and FL optimizes distributions induced by those policies rather than a common periodic acquisition process. Let θ denote shared convolutional and classifier parameters and φ_i the client-local BatchNorm state. Training minimizes

$$\min_{\theta,\,\{\phi_i\}_{i=1}^{N}} \sum_{i=1}^{N} \frac{n_i}{\sum_{j=1}^{N} n_j} \mathcal{L}(\widetilde{D}_i^{\pi_i};\theta,\phi_i). \quad (2)$$

Shared convolutional and classifier state is weighted at the server by each client's number of local training windows. Batch-normalization affine parameters φ_i and running statistics remain local following FedBN [8]. Only shared model state is used for training aggregation; raw signals, normalization statistics, and event timestamps remain local. For the reported audit, each client also returns its scalar threshold and aggregate evaluation metrics so that retention and threshold-group performance can be summarized. This is data locality, not a transport-security or formal privacy guarantee.

B. Send-on-Delta Policy

For normalized sample $x_{iw}(t)$ and last transmitted vector $v_{iw}(t^-)$, set $e_{iw}(0)=1$ so that the first sample is retained. For $t>0$, SOD retains the full three-axis vector when

$$e_{iw}(t) = \mathbf{1}[\|x_{iw}(t) - v_{iw}(t^-)\|_\infty \geq \delta_i]. \quad (3)$$

The infinity norm is taken across the x, y, and z axes, so an event occurs when any axis reaches the threshold. After an event, v_{iw} becomes $x_{iw}(t)$; otherwise, causal zero-order hold (ZOH) repeats the last transmitted vector on the original timeline. The retained count is $K_{iw}=\sum_{t=0}^{T-1} e_{iw}(t)$ and retention is $r_{iw}=K_{iw}/T$. Sending the full vector avoids three independent axis clocks and gives every method the same packet definition. In this evaluation, v_{iw} resets at each overlapping window; the experiment therefore isolates window-level acquisition and is not a trace of a continuously stateful physical sampler.

C. Exact Retained-Count Protocol

The confirmatory comparison must vary timestamp policy while holding packet count and downstream processing fixed. Periodic-matched distributes K_{iw} indices evenly and includes index zero. Random-matched includes index zero and selects K_{iw}-1 remaining indices uniformly without replacement using a deterministic preprocessing seed. All sparse streams use causal ZOH and satisfy

$$K_{iw}^{\text{periodic}} = K_{iw}^{\text{random}} = K_{iw}^{\text{SOD}} \quad \forall i, w. \quad (4)$$

Algorithm 1 is the exact protocol. The complete periodic stream is a high-cost reference, not a matched competitor. Linear interpolation was excluded because a value between two events requires a future event and is therefore noncausal for online classification; learned reconstruction would add a second model and compute confounder.

ALGORITHM 1. EXACT RETAINED-COUNT ACQUISITION PROTOCOL

Step	Operation
1)	Apply SOD to normalized window Xiw, always retain t=0, and obtain its retained index set and packet count Kiw.
2)	Select Kiw evenly spaced periodic packet indices, including zero.
3)	Select zero plus Kiw-1 seeded random packet indices without replacement.
4)	Reconstruct the three sparse streams on the original timeline using the same causal ZOH operator.
5)	Train and evaluate identical client partitions, model, optimizer, and seed for the three modes; report Kiw/T.

The random seed combines client, partition, mode, and window index, so masks are deterministic without being reused across subjects or windows. Prepared acquisition arrays are then fixed across optimization seeds. An exhaustive array audit found zero matched-count discrepancies across all 13,300 WISDM, 11,972 PAMAP2, and 20,453 MotionSense train, validation, and test windows.

In summary, Equations (1)-(3) realize C1 by linking each client's acquisition and reconstruction policy to its local learning distribution. Equation (4) and Algorithm 1 realize C2 by equalizing packet count and downstream processing while varying retained timestamps. The zero-mismatch audit establishes protocol conformance; recognition consequences are evaluated separately in Section V.

IV. EXPERIMENTAL METHOD

This section operationalizes the formulation and protocol without using test results to select the endpoint. It specifies the three client populations, leakage controls, shared model, paired-seed design, and quality gates needed to evaluate R1-R5 and obtain auditable evidence for F1-F3.

A. Datasets and Leakage Prevention

WISDM contains phone accelerometer recordings for six everyday activities [13]. All 36 subjects present in the downloaded WISDM v1.1 raw file are clients. PAMAP2 contains wearable measurements from 18 activities and nine participants [14]. We use the protocol chest 16-g accelerometer and retain the eight most represented activities. Eight subjects contribute to those activities. MotionSense contains 50-Hz iPhone recordings from 24 participants performing six activities in 15 trials [15]. We reconstruct three-axis device acceleration in units of g by adding the provided gravity and user-acceleration components. Each participant is one client.

Continuous subject-activity sequences are identified before splitting; a subject/activity change, timestamp reversal, or gap over 1 s for WISDM and 0.2 s for PAMAP2 starts a new sequence. Each MotionSense subject-trial file is preserved as a separate sequence. Every sequence is divided chronologically into 70% training, 15% validation, and 15% test segments before overlapping windows are generated. Consequently, no raw timestamp or overlapping sample crosses partitions. Sequences too short to supply all three partitions are logged and excluded. Each client's three-axis mean and standard deviation are fitted only on its eligible periodic training rows before acquisition encoding and then applied unchanged to validation and test data. Thresholds are expressed in those local standardized units and assigned cyclically across clients. MotionSense thresholds were fixed

from a label-free scan of training retention before any model result was inspected.

WISDM parsing accepted 1,098,207 labeled rows and logged one malformed record. Continuity segmentation excluded 139,485 rows in 14,992 short sequences that could not supply all three windowed partitions. The final preparation produced 9,766 training, 1,764 validation, and 1,770 test windows across six activities and all 36 clients. PAMAP2 parsing examined 2,872,533 protocol rows; we dropped 929,661 transient or unlabeled rows, 2,420 rows with missing selected sensor values, and 613 rows in two short selected-activity sequences. We retained cycling, ironing, lying, Nordic walking, sitting, standing, vacuum cleaning, and walking. Subject 109 contributed 6,391 labeled rows, all for rope jumping, so it had no rows in the selected set; eight clients remained. MotionSense parsing accepted all 1,412,865 rows from 360 subject-trial files with no malformed records; one 377-row subject-activity trial was too short for all three windowed partitions. It yielded 14,889 training, 2,780 validation, and 2,784 test windows across all 24 clients. Machine-readable metadata stores every exclusion and per-client class count.

TABLE I. DATASETS AND PREPROCESSING

Dataset	WISDM	PAMAP2	MotionSense
Clients	36	8	24
Classes	6	8	6
Sensor	Phone accel.	Chest accel.	Phone accel.
Window / overlap	128 / 50%	256 / 50%	128 / 50%
Train / val. / test	9766 / 1764 / 1770	8490 / 1741 / 1741	14889 / 2780 / 2784
Thresholds	.40, .80, 1.20, 2.00	.05, .08, .10, .15	.05, .10, .20, .40
Test sparse retention	50.5%	57.4%	46.1%

Because standardization uses $z_{i,c}=(x_{i,c}-\mu_{i,c})/\sigma_{i,c}$, a standardized threshold maps to an axis-specific raw-unit change of $\delta_i\sigma_{i,c}$. The lowest and highest groups span .904-4.432 and 6.278-16.062 WISDM raw units (10 approximately equals 1 *g*), .075-.254 and .196-.782 m/s^2 on PAMAP2, and .0176-.0375 and .109-.274 *g* on MotionSense. These ranges quantify why the policies are not identical physical or energy budgets. The supplementary calibration table reports every client-axis value.

B. Model and Federated Training

The compact one-dimensional convolutional neural network (CNN) has widths 16, 32, and 32; kernel sizes 5, 5, and 3; rectified linear units; two max-pooling stages; adaptive average pooling; and a linear classifier. Batch normalization follows the first two convolutions. PyTorch implements optimization [16], and Flower NumPy clients provide the federated client interface [17]. The seeded in-process orchestrator uses full client participation, 30 communication rounds, one local epoch, Adam with learning rate 0.001, batch size 64, and aggregation weighted by local training-window count. A fresh Adam optimizer is instantiated for each participating client in each round, so optimizer state is neither shared nor retained across rounds. FedBN excludes all batch-normalization state from server aggregation.

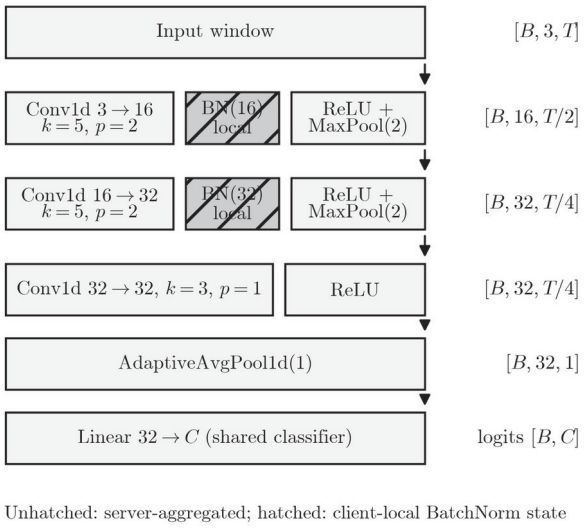


Fig. 1. Implemented CNN and FedBN parameter ownership. Convolutional and classifier parameters are aggregated by local training-window count; both BatchNorm affine parameters and running statistics remain client-local. *T=128,C=6* for WISDM and MotionSense; *T=256,C=8* for PAMAP2.

The three-channel network contains 6,246 trainable parameters for the six-class datasets and 6,312 for PAMAP2; only the output layer changes with class count. BatchNorm contributes 96 local trainable parameters. The shared parameter counts are 6,150 and 6,216, respectively. This deliberately small architecture limits tuning burden and is compatible with an edge-oriented study, although no hardware latency is claimed. Full client participation removes variability from client availability, allowing each seed to pair acquisition modes over the same federation.

A predeclared capacity sensitivity doubles every convolutional width, producing widths 32, 64, and 64. The enlarged network has 23,750 trainable parameters (23,558 shared) for six classes and 23,880 (23,688 shared) for eight classes, about 3.8 times as many trainable parameters as the primary model. It changes no acquisition arrays, rounds, optimizer settings, endpoints, or seeds and compares only SOD with periodic-matched sampling on WISDM and PAMAP2.

All four confirmatory acquisition modes use the same model width and three input channels: full periodic, periodic matched, random matched, and SOD with ZOH. Development experiments considered event masks, elapsed time since the previous event, normalized jumps, and threshold metadata. Those development representations used additional input channels, did not generalize across the original two datasets, and are not presented as the main method.

C. Confirmatory Analysis

The initial optimization seed set {7, 19, 41, 73, 101, 137, 173, 211} was fixed before the original confirmatory batch. Before any extension result was inspected, a complete review-response plan fixed four additional seeds {251, 293, 337, 383} for every mode and the original two datasets; all planned runs are retained. MotionSense was then predeclared as a complete twelve-seed, four-mode external-validation block before any model result was inspected. Acquisition arrays, including random-retention masks, are fixed across the twelve seeds, so intervals quantify optimization variability conditional on one

prepared dataset. The confirmatory client loop trains only on training windows; prepared validation windows are not used to select a checkpoint or the round-30 endpoint. The endpoint is the test-sample-weighted mean of per-client Macro-F1 scores at round 30, not the best test round. Each client is scored over the global class set, and its Macro-F1 is weighted by its number of test windows; this is not class-frequency-weighted F1. SOD is paired with periodic-matched and random-matched by optimization seed. We report the mean paired difference, a two-sided 95% t confidence interval, paired Cohen's d_z, a paired t test, and an exact two-sided Wilcoxon signed-rank test. Holm's procedure controls family-wise error across the two primary comparisons within each dataset [18]. The t-based Holm-adjusted p values determine confirmatory significance; Wilcoxon results are sensitivity checks. No seed or completed run was removed.

Selecting a fixed final round prevents the test set from choosing a favorable checkpoint. Pairing by seed removes between-seed variability from each method difference. The paired t interval summarizes effect magnitude; the exact Wilcoxon test provides a rank-based sensitivity that does not rely on normality of paired differences. Statistical tests are applied only to the two declared packet-count-matched comparisons, not to every development ablation.

D. Reproducibility and Quality Gates

The confirmatory environment used Python 3.12.13, NumPy 2.5.1, SciPy 1.18.0, PyTorch 2.13.0, and Flower 1.32.1. Configuration and metadata files define windows, thresholds, clients, rounds, and optimizer settings. Each confirmatory CSV contains 30 round rows, retention, and both sample-weighted and unweighted client metrics; its filename records the seed. Focused group-check reruns additionally contain threshold-group metrics. Analysis scripts read only the final row for confirmatory reporting and generate tables and figures from CSV files.

All 37 automated tests passed. Fourteen cover acquisition, reconstruction, exact matching, and sequence-state reconstruction; twelve cover parsing, preprocessing, chronological splitting, training-only normalization, and window isolation; nine cover the model, metrics, federated aggregation, imports, and experiment plumbing; and two verify augmentation-budget invariants. They establish software conformance to Algorithm 1 and the leakage constraints, not application accuracy. Recognition evidence instead comes from the 144 complete primary mode-dataset-seed files and the numerical audit that independently reconstructs Tables II and III. A reviewer-facing table reports all 144 seed-level endpoints. The workspace was not initialized as a Git repository, so a commit identifier cannot be reported. The supplementary archive contains one deduplicated raw round-level result table; source, configurations, commands, and the environment record are retained by the author.

In summary, the method fixes acquisition arrays, partitions, optimization schedules, and a round-30 endpoint before comparison, while pairing acquisition modes by seed. The 37 tests verify the protocol, reconstruction, leakage controls, and aggregation behavior; they do not substitute for the recognition estimates and uncertainty reported next.

V. Results

This section tests the three empirical findings in the order declared in Section I-B. The primary matched-cost analysis evaluates F1 and F2, while threshold-group, policy-shift, sequence-stateful, capacity, and augmentation analyses establish the scope and failure conditions summarized by F3.

A. Matched-Cost Performance

TABLE II. ROUND-30 RECOGNITION AND RETENTION, TWELVE-SEED MEANS

Data	Acquisition	Macro -F1	Acc.	Macro-F1 95% CI	Ret.
WISDM	Full periodic	.5448	.8656	[.5138, .5759]	1.000
WISDM	Periodic matched	.4992	.8401	[.4667, .5317]	.505
WISDM	Random matched	.4745	.8274	[.4385, .5106]	.505
WISDM	SOD ZOH	.5113	.8450	[.4777, .5448]	.505
PAMAP2	Full periodic	.7090	.7487	[.6990, .7190]	1.000
PAMAP2	Periodic matched	.7068	.7464	[.6979, .7156]	.574
PAMAP2	Random matched	.7092	.7486	[.6988, .7197]	.574
PAMAP2	SOD ZOH	.7077	.7469	[.6977, .7177]	.574
MotionSense	Full periodic	.8451	.9125	[.8387, .8516]	1.000
MotionSense	Periodic matched	.8439	.9126	[.8365, .8514]	.461
MotionSense	Random matched	.8425	.9107	[.8346, .8504]	.461
MotionSense	SOD ZOH	.8450	.9125	[.8382, .8518]	.461

Table II and Fig. 2 show different regimes. Macro-F1 is intentionally conservative: it is computed separately for each client over the global class set, including activities absent from that client's test partition, and only then weighted by client test-window count. It is therefore lower than pooled or class-frequency-weighted scores. Accuracy is reported to make absolute recognition behavior visible, but paired Macro-F1 remains the confirmatory endpoint because it gives every activity equal weight within each client.

On WISDM, SOD yields Macro-F1 .5113 and accuracy .8450 while retaining 50.5% of the full-periodic packet count. The paired difference over periodic matched is .0120 (95% CI [.0054, .0187]). Its effect size is d_z=1.16; Holm p=*.0041* and exact Wilcoxon p=*.0015*. The paired difference over random matched is .0368 (95% CI [.0105, .0630]). Its effect size is d_z=.89; Holm p=*.0104* and Wilcoxon p=*.0005*. Because within-window retained packet counts, reconstruction, architecture, training schedule, and seeds are matched, this establishes finding F1: signal-dependent timestamp selection produces higher WISDM Macro-F1 than the two controls.

The direction is also consistent at the seed level. SOD exceeds periodic matched in eleven of twelve WISDM seeds and random matched in all twelve. The only negative SOD-periodic difference is -.0020, whereas positive differences range from .0010 to .0375. SOD-random gains range from .0018 to .1573. The latter interval still reflects one large seed effect, but the exact signed-rank result and twelve consistent directions show that the conclusion is not produced by that seed alone. One added seed produced low absolute scores for every mode and was retained, widening the marginal confidence intervals without reversing the paired effects.

On PAMAP2, the SOD-periodic difference is .0009 (95% CI [-.0015, .0033], Holm p=*.7393*), and the SOD-random difference is -.0016 (95% CI [-.0052, .0021], Holm p=*.7393*). MotionSense supplies a higher-performing third condition: SOD reaches Macro-F1 .8450 and accuracy .9125 at 46.1% retention. Its SOD-periodic difference is .0011 (95% CI [-.0018, .0040], Holm p=*.4305*), and SOD-random is .0025 ([-.0007, .0057],

Holm $p=.2246$). Full and sparse observed means remain close on both datasets despite 42.6% and 53.9% fewer retained three-axis packets, respectively. These tests do not establish statistical equivalence because no equivalence margin was predeclared. They establish finding F2 as a replicated scope boundary: sparse-ZOH and full-input observed means can be close, but there is no evidence that SOD is universally more accurate than packet-count-matched sampling.

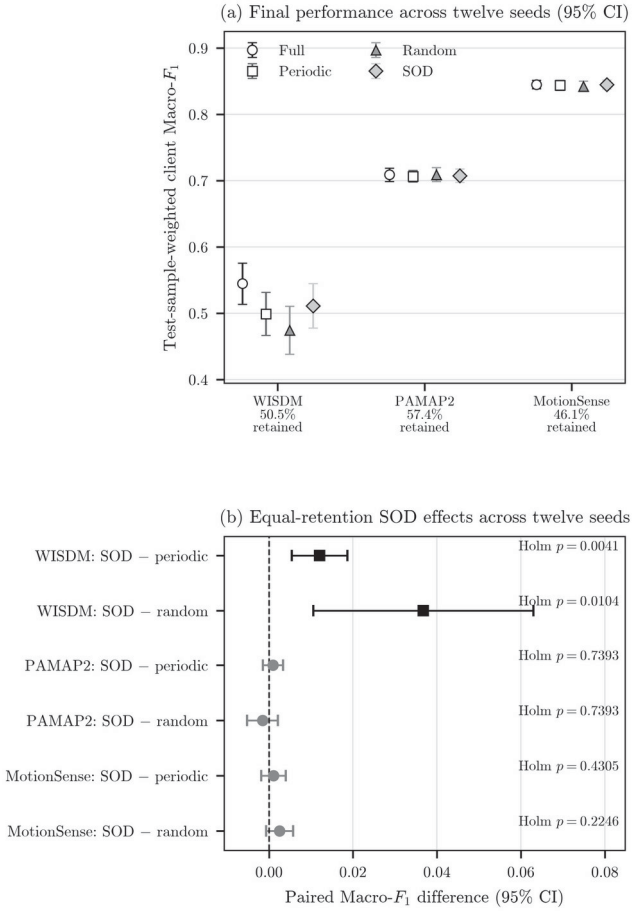


Fig. 2. Cross-dataset results generated from the locked twelve-seed tables. (a) Round-30 test-sample-weighted mean client Macro-F1 with two-sided 95% t confidence intervals. (b) Paired Send-on-Delta effects at equal retention; black squares have Holm-adjusted $p<.05$ and gray circles do not.

TABLE III. PAIRED SOD MACRO-F1 EFFECTS AT EQUAL RETENTION

Data	Baseline	Macro-F1 diff.	95% CI	dz	Holm p
WISDM	Periodic	.0120	[.0054,.0187]	1.16	.0041
WISDM	Random	.0368	[.0105,.0630]	.89	.0104
PAMAP2	Periodic	.0009	[-.0015,.0033]	.25	.7393
PAMAP2	Random	-.0016	[-.0052,.0021]	-.27	.7393
MotionSense	Periodic	.0011	[-.0018,.0040]	.24	.4305
MotionSense	Random	.0025	[-.0007,.0057]	.50	.2246

B. Acquisition Heterogeneity and Policy Shift

Client-group reporting reveals a cost-performance fairness issue hidden by federation-level means. In the eight-seed policy audit, the WISDM training-policy mean across seed-level weakest threshold groups was .4447 (95% CI [.4087, .4806]), compared with federation-level Macro-F1 .5156. Under the severe policy, the weakest-group mean fell to .2227 (95% CI [.2061, .2393]). PAMAP2 weakest-group means remained between .6490 and .6576 across all three policies. Threshold-conditioned performance should therefore be monitored rather than inferred from retention alone.

These group summaries are diagnostic, not randomized threshold-dose estimates. Thresholds were assigned cyclically, but groups contain different subjects and window counts, while client-local standardization makes each threshold relative to that client's training variability. A group difference can therefore reflect acquisition severity, subject difficulty, activity mix, or their interaction. The confirmatory policy comparisons instead come from modes evaluated on the same client windows and paired by seed.

The group audit separates average efficiency from service reliability. The federation-level mean reports overall utility, whereas the lowest group indicates whether a sensing-budget cohort is poorly served. A policy can improve the mean at matched cost and still create a tail-risk problem. Reporting both quantities prevents the aggregate from hiding clients whose local streams have become too sparse and supplies a criterion for fall-back thresholds or per-client recalibration.

For the policy-shift audit, training arrays remain identical to the main experiment and only test encodings change. The same eight optimization seeds evaluate training, moderate unseen, and severe threshold policies on each dataset. These are exploratory paired sensitivity estimates rather than additional confirmatory hypothesis tests.

TABLE IV. EIGHT-SEED SOD POLICY-SHIFT AUDIT

Data	Evaluation policy	Ret.	SOD Macro-F1	Change	Worst-group Macro-F1
WISDM	Training	.505	.5156	--	.4447
WISDM	Moderate	.417	.5096	-.0060	.4159
WISDM	Severe	.277	.4487	-.0669	.2227
PAMAP2	Training	.574	.7117	--	.6497
PAMAP2	Moderate	.496	.7094	-.0023	.6490
PAMAP2	Severe	.355	.7118	+.0001	.6576

On WISDM, the moderate policy lowers retention by 8.8 percentage points and changes Macro-F1 by -.0060 (95% paired CI [-.0106, -.0013]). The severe policy lowers retention by 22.8 points and changes Macro-F1 by -.0669 (95% paired CI [-.0804, -.0534]); its weakest-group mean is .2227. PAMAP2 behaves differently: even the severe policy lowers retention by 21.9 points while the paired Macro-F1 change is +.0001 (95% CI [-.0031, .0034]) and the weakest-group mean is .6576. The observed failure boundary therefore depends on the interaction between aggressive sparsity and the activity stream, not on threshold magnitude alone.

An exploratory sequence-stateful sensitivity probes the primary protocol's per-window reset. Full-input windows were stitched only where their overlaps matched exactly; acquisition state then persisted within each inferred sequence and reset at sequence or partition boundaries. SOD, periodic, and random policies retained exactly the same packet count per reconstructed sequence. Across eight fixed seeds, WISDM trace retention was 49.7%; SOD-periodic changed by +.0115 (95% CI [-.0017, .0247]) and SOD-random by +.0557 ([.0134, .0979]). At 56.9% PAMAP2 retention, the respective changes were +.0027 ([-.0007, .0060]) and +.0040 ([-.0032, .0111]). At 47.8% MotionSense retention, SOD-periodic was +.0022 ([-.0010, .0055]) and SOD-random was +.0056 ([.0012, .0101]). Thus, SOD exceeds random retention in the stateful WISDM and MotionSense conditions, but every stateful SOD-periodic inter-

val crosses zero. This sensitivity narrows F1 to the declared window-level protocol and does not substitute for continuously stateful hardware collection.

The predeclared capacity sensitivity tests whether the primary result is peculiar to the compact CNN. With doubled channel widths and about 3.8 times as many trainable parameters, WISDM SOD reaches .6311 versus .6142 for periodic matched, a +.0169 difference (95% CI [.0120, .0219], paired $p<.0001$). PAMAP2 reaches .7216 versus .7186, a smaller +.0030 difference ([.0006, .0055], paired $p=.0204$). The WISDM direction therefore persists with increased capacity; the larger model produces a statistically detectable but practically small PAMAP2 difference. Because this sensitivity covers one baseline, one architecture family, and the original two datasets, it supports robustness to width rather than a model-independent claim.

C. Negative Augmentation Audit

Threshold augmentation was examined as a possible policy-shift remedy. An initial implementation re-encoded every training window at four thresholds, but this also quadrupled the number of optimization examples and was therefore confounded. It was not used for any paper claim. A corrected compute-matched version selected one augmented threshold per original example. It reduced WISDM main-policy Macro-F1 by .0619 (95% paired CI [-.0851, -.0386]); its PAMAP2 main-policy difference was +.0020 with a 95% confidence interval that crossed zero. Together with the weakest-group policy-shift result, this establishes finding F3: the tested threshold-randomization scheme can erase useful acquisition structure, while aggressive sparsity needs targeted adaptation or a fallback policy.

In summary, F1 is supported on WISDM under the declared window-level protocol, whereas PAMAP2 and MotionSense establish F2 by bounding any claim of universal advantage. The policy-shift, weakest-group, sequence-stateful, capacity, and augmentation results establish F3 as a deployment warning: mean matched-cost performance does not guarantee group reliability or robustness to a changed sensing policy.

VI. Discussion

This section interprets C1-C2 and F1-F3 against the application requirements stated before the method. It distinguishes contributions to scientific knowledge from implementation support, translates the estimates into deployment guidance, and states the limits that prevent broader energy, hardware, or architecture-independent claims.

A. Interpretation

The WISDM effect supports the paper's main systems claim: retained packet count alone does not fully characterize sparse federated sensing. SOD selects threshold-crossing timestamps, whereas periodic and random baselines use the same retained packet count without conditioning on signal change. PAMAP2 and MotionSense provide no primary packet-count-matched evidence of an SOD advantage. Their full and sparse observed means are close at the tested thresholds, but the reported tests are not equivalence tests. A realistic conclusion is therefore conditional: signal-dependent sampling can improve evidence efficiency, but the value depends on sensor placement, activity dynamics, threshold range, task redundancy, and model capacity.

A simple delta trigger can also fire on vibration or sensor noise. The WISDM result therefore does not show that every SOD event is semantically meaningful or that difference-only triggering is universally appropriate; it shows only that the selected timestamps produce higher Macro-F1 than the two packet-count-matched controls in this prepared condition. Hysteresis, multi-sample change tests, or learned event criteria may be needed for noisier deployments and should be evaluated under the same exact packet-count protocol.

FedBN is an implementation choice rather than a new contribution. Development pilots showed that aggregating batch-normalization state harmed sparse heterogeneous inputs; keeping it local improved all tested modes. A three-seed interaction analysis found no SOD-specific FedBN gain, so normalization is not claimed as the mechanism behind the packet-count-matched effect.

The full-periodic comparison also differs across datasets. WISDM loses .0335 Macro-F1 when replacing full input with SOD at roughly half retention, whereas PAMAP2 and MotionSense observed means differ by .0013 and .0001. Long PAMAP2 windows, chest placement, sustained activities, and MotionSense trial regularity may provide enough temporal redundancy for ZOH at the tested thresholds; more aggressive thresholds or transition-focused labels could produce a different regime.

The enlarged-CNN sensitivity further shows that the WISDM SOD-periodic direction is not confined to the primary parameter count. Its small positive PAMAP2 effect also cautions against treating an interval crossing zero from one architecture as proof that acquisition policy is irrelevant. Model capacity is therefore another boundary condition, although the unchanged architecture family does not establish generality across temporal models.

B. Application-Requirement Assessment

Table V evaluates requirements R1-R5 from Section I-A. R1 is satisfied by construction and exhaustive count checking. R2-R4 are satisfied as evaluation requirements, but the measured utility is dataset-dependent and the group and policy-shift values expose deployment limitations rather than universal service reliability. R5 is only partially demonstrated because parameter count and raw-signal locality are known, while hardware latency, radio traffic, and energy were not measured.

TABLE V. APPLICATION-REQUIREMENT ASSESSMENT

Req.	Quantitative evidence	Assessment
R1	45,725 windows; zero count mismatches	Met
R2	WISDM CIs exclude zero; other primary CIs cross zero	Met; benefit dataset dependent
R3	WISDM weakest-group mean .4447; severe .2227	Met as evaluation; no floor
R4	WISDM change -.0669; PAMAP2 +.0001	Met; boundary is dataset dependent
R5	6,150 / 6,216 shared; 2x width, about 3.8x parameters; no hardware test	Partial

The 37 automated tests support R1 and implementation validity by detecting count, reconstruction, leakage, and aggregation failures. They cannot establish R2-R5 by themselves. Those requirements need recognition results, group metrics, policy-shift evaluation, model-size accounting, and eventually

hardware measurements. This distinction places the software artifact in its proper role: it makes the empirical knowledge auditable but is not itself the scientific finding.

C. Practical Implications

Deployment should use a task-specific packet-count-matched curve rather than a universal threshold, report both federation-level and group outcomes, validate battery-driven threshold changes as data-policy shifts, and retain a fallback threshold when group performance falls below an application requirement.

Calibration records should be versioned by sensor placement and client population, including training-only normalization, thresholds, retention curves, mean and worst-group Macro-F1, and seed uncertainty. Firmware, mounting, or population changes should trigger revalidation.

The window-level retention metric is the fraction of original timestamps at which a three-axis packet is retained; it is not the classifier tensor size or a direct deployment traffic trace. Real radios and continuously stateful samplers introduce cross-window history, wake-up, header, retransmission, batching, and idle costs, so retention cannot be converted directly into traffic or battery-life improvement. A hardware study should measure those components and may find that sending bursts or locally buffering events is more efficient than transmitting each event immediately.

The negative augmentation result shows that the tested compute-matched per-window threshold randomization is not a reliable robustness strategy in this setting. Future work could constrain changes to nearby thresholds, distill from dense input, or optimize worst-threshold performance under a packet budget; such tests should remain compute matched and predeclare policy-shift endpoints.

D. Threats to Validity

The study has five main limitations. First and most importantly, the primary SOD streams are retrospectively generated and reset at each overlapping window. Events between original timestamps cannot be recovered, and primary retention is a window-level protocol metric rather than a physical packet trace. The eight-seed sequence-stateful sensitivity preserves state across inferred prepared sequences, but it is still retrospective and resets at sequence and partition boundaries. Neither protocol supports measured radio-traffic or battery-savings claims. Second, WISDM continuity filtering removed 12.7% of parsed rows in short fragments and may alter client or activity representation. Third, thresholds are fixed standardized values rather than calibrated from hardware noise or an energy model. The reported client-axis conversion quantifies their physical scale but does not turn them into equal hardware budgets. Fourth, the eight-seed policy-shift and threshold-group audits remain exploratory because their endpoints were not predeclared as primary, and threshold groups contain different subjects rather than randomized threshold assignments. Fifth, only accelerometers and ZOH reconstruction are considered. Linear interpolation is noncausal if it uses a future event, while learned reconstruction would add model and compute confounders; ZOH was selected as the simplest causal common representation.

There are also statistical limitations. Twelve paired optimization seeds improve primary effect precision, but uncertainty remains, especially for the WISDM random baseline. Because acquisition arrays are fixed, the intervals do not include variabil-ity from resampling random-retention masks or client thresholds. The policy-shift and sequence-stateful analyses use eight seeds. PAMAP2 and MotionSense primary intervals crossing zero do not establish equivalence. The capacity sensitivity varies width within the same CNN family and evaluates only SOD against periodic retention; it cannot establish architecture-independent behavior. Finally, the three public datasets differ in more than sampling policy, so they reveal external-validity boundaries but do not isolate which dataset characteristic causes the interaction.

In summary, R1-R4 are satisfied as evaluation requirements, while R5 remains partial because no physical traffic, latency, or energy measurement was made. The scientific contribution is therefore conditional knowledge about acquisition-policy heterogeneity and timestamp selection, supported by a controlled protocol and bounded by cross-dataset, stateful, capacity, and policy-shift evidence. The implementation makes that knowledge auditable but is not itself the scientific contribution.

VII. Conclusion

This study makes two formal contributions. C1, represented by Equations (1)-(3), defines acquisition-policy heterogeneity by showing how client-specific acquisition, reconstruction, and threshold choices induce the local datasets optimized by FL. C2, represented by Equation (4) and Algorithm 1, isolates timestamp policy by holding within-window packet count, causal reconstruction, client partitions, model, optimizer, and seed fixed. The audit found zero retained-count mismatches across 45,725 prepared windows. These contributions provide a referable formulation and controlled evaluation protocol; they are not presented as a new FL algorithm or a new SOD sampler.

The protocol yields three empirical findings. F1 is a positive but bounded WISDM result: at 50.5% retention, SOD exceeds periodic matched by .0120 (95% CI [.0054, .0187], Holm $p=.0041$) and random matched by .0368 ([.0105, .0630], Holm $p=.0104$). The SOD-periodic direction also persists with doubled convolutional widths at +.0169 ([.0120, .0219]). F2 is the external-validity boundary: PAMAP2 differences are +.0009 against periodic and -.0016 against random, while MotionSense differences are +.0011 and +.0025; all four primary confidence intervals cross zero. In addition, every sequence-stateful SOD-periodic interval crosses zero, so the WISDM finding is tied to the declared acquisition protocol and does not establish universal superiority or equivalence elsewhere.

F3 identifies a deployment boundary that federation-level means can hide. Across eight policy-shift seeds, severe WISDM sparsity changes Macro-F1 by -.0669 (95% CI [-.0804, -.0534]) and lowers the weakest-group mean to .2227. Under the PAMAP2 severe policy, the mean change is +.0001 ([-.0031, .0034]). The tested compute-matched threshold randomization is not a remedy: it changes WISDM main-policy Macro-F1 by -.0619 ([-.0851, -.0386]).

The contribution to scientific knowledge is therefore the demonstrated need to treat acquisition policy as a pretraining source of client heterogeneity whose effect depends on the dataset, protocol, model, and deployment policy. For practice, the results support task-specific packet-count-matched curves, threshold-group monitoring, and revalidation after sensing-policy changes. They do not establish hardware energy savings, continuously stateful physical-sampler behavior, or suitability for safety-critical HAR.

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS), 2017, pp. 1273-1282.

[2] C. Li, D. Niu, B. Jiang, X. Zuo, and J. Yang, "Meta-HAR: Federated representation learning for human activity recognition," in Proc. Web Conf. (WWW), 2021, pp. 912-922, doi: 10.1145/3442381.3450006.

[3] M. Miśkowicz, "Send-on-Delta concept: An event-based data reporting strategy," Sensors, vol. 6, no. 1, pp. 49-63, 2006, doi: 10.3390/s6010049.

[4] M. Díaz-Cacho, E. Delgado, A. Barreiro, and P. Falcón, "Basic Send-on-Delta sampling for signal tracking-error reduction," Sensors, vol. 17, no. 2, Art. no. 312, 2017, doi: 10.3390/s17020312.

[5] M. Lewandowski, B. Płaczek, and M. Bernas, "Classifier-based data transmission reduction in wearable sensor network for human activity monitoring," Sensors, vol. 21, no. 1, Art. no. 85, 2021, doi: 10.3390/s21010085.

[6] G. Yang, L. Zhang, C. Bu, S. Wang, H. Wu, and A. Song, "FreqSense: Adaptive sampling rates for sensor-based human activity recognition under tunable computational budgets," IEEE J. Biomed. Health Inform., vol. 27, no. 12, pp. 5791-5802, 2023, doi: 10.1109/JBHI.2023.3321639.

[7] M. Pabian, D. Rzepka, M. Pawlak, M. Miśkowicz, and R. Sroka, "Signal-to-event encoding parameter selection for multiple event classification with spiking neural networks," Front. Neurosci., vol. 19, Art. no. 1610766, 2025, doi: 10.3389/fnins.2025.1610766.

[8] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," in Proc. Int. Conf. Learn. Representations (ICLR), 2021.

[9] D. Fenoglio et al., "Multi-frequency federated learning for human activity recognition using head-worn sensors," in Proc. Int. Conf. Intell. Environ. (IE), 2024, pp. 17-24, doi: 10.1109/IE61493.2024.10599924.

[10] S. Karilanova, S. Dey, and A. Özçelikkale, "Federated learning of spiking neural networks under heterogeneous temporal resolutions," arXiv:2605.15355, 2026.

[11] N. Skatchkovsky, H. Jang, and O. Simeone, "Federated neuromorphic learning of spiking neural networks for low-power edge intelligence," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2020, pp. 8524-8528, doi: 10.1109/ICASSP40776.2020.9053861.

[12] S. Zehtabi, S. Hosseinalipour, and C. G. Brinton, "Decentralized event-triggered federated learning with heterogeneous communication thresholds," in Proc. 61st IEEE Conf. Decis. Control (CDC), 2022, pp. 4680-4687, doi: 10.1109/CDC51059.2022.9993258.

[13] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," ACM SIGKDD Explor. Newsl., vol. 12, no. 2, pp. 74-82, 2011, doi: 10.1145/1964897.1964918.

[14] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in Proc. 16th IEEE Int. Symp. Wearable Comput. (ISWC), 2012, pp. 108-109, doi: 10.1109/ISWC.2012.13.

[15] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Mobile sensor data anonymization," in Proc. Int. Conf. Internet Things Design Implement. (IoTDI), 2019, pp. 49-58, doi: 10.1145/3302505.3310068.

[16] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 8024-8035.

[17] D. J. Beutel et al., "Flower: A friendly federated learning research framework," arXiv:2007.14390, 2020.

[18] S. Holm, "A simple sequentially rejective multiple test procedure," Scand. J. Stat., vol. 6, no. 2, pp. 65-70, 1979, doi: 10.2307/4615733.