

A Multi-Radio Audit of Subject-Dependent Performance in Recent Public WiFi CSI Datasets

Apil Khadka

Nepal College of Information Technology

Lalitpur, Nepal

apil.231210@ncit.edu.np

Sanjana Paudel

Nepal College of Information Technology

Lalitpur, Nepal

sanjana.231641@ncit.edu.np

Abstract—Random train/test splits of WiFi channel state information (CSI) sensing data can place the same subjects and overlapping temporal windows in both partitions, allowing subject- or session-specific signal structure to inflate estimates of activity recognition performance. We conduct a multi-radio evaluation audit of WiStride, ElderAL-CSI, and the activity-recognition subset of SHARP-80 using random forest and neural-network probes under random-window and subject-disjoint protocols, with majority baselines and macro-averaged F1 scores. On WiStride, random forest accuracy decreases from 0.773 to 0.684, a gap of 8.9 percentage points (subject-cluster bootstrap 95% confidence interval: 6.4–12.1 percentage points), while macro-F1 decreases from 0.526 to 0.322. SHARP-80 shows a larger leave-one-subject-out collapse under campaign-linked person labels, with RF accuracy falling from 0.954 to 0.181. Its four groups and ElderAL’s three subjects provide preliminary evidence. Five repeated source-group seeds show that cross-radio walk/non-walk transfer does not exceed target majority accuracy. A group-ID predictability diagnostic on WiStride exceeds its majority baseline, indicating subject/session-correlated structure without isolating its causal source. These results motivate subject-disjoint evaluation, majority baselines, and imbalance-aware metrics in CSI reporting, and we provide a reporting checklist and reproducibility artifacts to facilitate their use.

Keywords—WiFi channel state information, human activity recognition, subject-disjoint evaluation, evaluation protocol, cross-dataset evaluation, data leakage

I. INTRODUCTION

Consider a public WiFi channel state information (CSI) gait corpus evaluated two ways on the same windows and with the same random forest (RF) probe. Under a stratified random train/test split, accuracy (Acc) looks strong. Under a subject-disjoint split, Acc drops by several percentage points and macro-averaged F1 (macro-F1) drops more sharply. This discrepancy reflects the evaluation protocol rather than a change in the underlying task. The optimistic split allowed the same people and overlapping temporal windows into both partitions, so person- or session-specific multipath could masquerade as activity recognition [1]–[4].

Contactless CSI sensing is otherwise attractive for human activity recognition (HAR), gait analysis, and aging-in-place monitoring [5], [6]. Public datasets and deep learning pipelines have proliferated, and many published numbers still come from random packet- or window-level splits. Subject-disjoint evaluation is a prerequisite for claims about activity generalization. Prior protocol critiques mainly reanalyzed established CSI-HAR benchmarks. Recent IEEE DataPort releases span

heterogeneous radios, including ESP32, commercial 802.11ax, and Nexmon 80 MHz platforms [7]–[10]. The empirical question is whether those recent corpora still show the same protocol gap, and whether amplitude-only recognition transfers across radio stacks once subjects are held out.

This paper addresses that question with a *multi-radio evaluation audit*, not a new recognizer. We ask:

How large is the protocol gap Δ in Acc under random versus subject-disjoint splits on recent public CSI datasets, and do simple baselines transfer across datasets under subject-aware protocols?

Our contributions are:

- 1) A unified amplitude-window pipeline that maps WiStride, ElderAL-CSI, and a SHARP-80 activity-recognition (AR) subset into comparable tensors with subject group IDs across three radio front ends (Table I, Fig. 1, and Section III).
- 2) A protocol-gap metric Δ (random Acc minus subject Acc) for RF, a one-dimensional convolutional neural network (CNN1D), and a convolutional neural network–gated recurrent unit (CNN–GRU), reported with means and standard deviations, majority baselines, macro-F1, and a subject-cluster uncertainty interval for WiStride RF Δ (Tables II and III, and Sections IV and V).
- 3) A purged loader-order group-ID RF probe that predicts subject/campaign IDs from the historical preprocessing population as a subject/session fingerprint diagnostic (Table V).
- 4) A cross-dataset *walk vs. non-walk* study with zero-shot RF, correlation alignment (CORAL), an in-dataset positive control, and a majority baseline on every target (Tables VI and VII).
- 5) A short reporting checklist for CSI papers and DataPort releases (Section VI-E), with the full recomputation workflow in the repository notebooks and CLI, current tables and figures tracked in the repository, and full run metadata retained in the controlled evidence packages.

Inferential priority throughout: WiStride (25 subjects) is primary evidence. SHARP-80 leave-one-subject-out shows a large decline across four campaign-linked person labels. Both SHARP-80 and ElderAL (3 subjects) provide preliminary evidence only (Section VI).

II. RELATED WORK

A. Evaluation Protocol and Subject Splits

In a series of critical reanalyses, Varga and colleagues showed that published CSI-HAR pipelines can drop dramatically when person-independent partitions replace random splits, and they recommend subject-based partitioning together with metrics beyond Acc alone [1]–[4]. Building on that foundation, we treat random Acc as an *optimistic estimate*, not evidence of activity generalization. Our contribution is to apply the same lens (majority baselines, macro-F1, a group-ID triangulation probe, and a shared code path) to *newer, multi-radio* IEEE DataPort releases.

B. WiFi CSI Sensing and Benchmarks

SenseFi provides a library and deep learning benchmark across multiple model families [5]. Surveys on WiFi sensing generalizability organize failure modes by environment, user, and device shift [6]. Our cross-radio walk/non-walk study sits at the intersection of user and device shift under subject-aware labels. Cross-domain gesture recognition with body-coordinate velocity profiles (BVP, Widar3.0) remains a canonical transfer setting [11]. Early deep CSI gait work similarly showed strong person-discriminative structure in CSI [12]. Classic Intel 5300 tooling [13], Nexmon-based CSI HAR pipelines [14], and newer commercial 802.11ax CSI extractors [10] further diversify the hardware landscape.

C. Datasets Used in This Study

WiStride provides ESP32 CSI from indoor walking subjects [7]. ElderAL-CSI targets aging-in-place activities with ZTE amplitude CSV exports [8], [10]. Meneghello and colleagues release an 80 MHz Nexmon CSI corpus used to validate SHARP [9], [15], [16]. We use a documented AR subset spanning four subjects (P1–P4) as our third hardware stack.

a) *Positioning*: Unlike SenseFi (a multi-model benchmark) or Widar3.0 (domain-invariant features), we do not propose a new recognizer or representation. We quantify protocol gaps and majority-controlled cross-radio transfer on recent heterogeneous DataPort corpora, an evaluation audit that remains less common than publishing new models.

III. DATASETS AND EVALUATION PROTOCOL

The evaluation problem is fixed before models are chosen: build comparable amplitude windows on three recent DataPort corpora, attach a subject group ID to every window, and measure the gap between random-window Acc and subject-disjoint Acc under identical features. Cross-radio transfer uses a shared binary label present on all three stacks. Caveats that affect interpretation (ElderAL’s three subjects, SHARP-80’s campaign-linked P1–P4 labels, and WiStride stop imbalance) are stated where they arise and revisited in Section VI.

TABLE I
DATASETS AFTER AMPLITUDE-WINDOW PREPROCESSING.

Dataset	Hardware	Shape (N, T, C)	Subjects
WiStride	ESP32 CSI	(3975, 100, 64)	25
ElderAL-CSI	ZTE AX3000	(2175, 16, 64)	3
SHARP-80 AR	Nexmon 80 MHz	(4760, 100, 64)	4

A. Datasets

Table I summarizes the three corpora after preprocessing. Fig. 1 shows window counts per label. We use *subject* uniformly for the grouping unit (WiStride person ID; ElderAL person ID; SHARP-80 P1–P4).

a) *WiStride*: WiStride contains ESP32 CSI (64 subcarriers) from 25 indoor walking subjects. Fine-grained event tags are coarsened to {walk, turn, stop} for a compact three-class taxonomy. Sliding windows use length $T=100$ and hop 50 (50% overlap), with a per-subject packet cap of 8000, yielding 3975 windows. Class stop covers 2848/3975 windows ($\approx 71.7\%$), so an always-stop classifier already reaches $\text{Acc} \approx 0.72$. Macro-F1 and a majority baseline are therefore reported alongside Acc.

b) *ElderAL-CSI*: ElderAL-CSI contains recordings from three subjects across multiple positions and six documented daily activities. ZTE amplitude columns (`amp_*`) are downsampled to 64 channels. Each CSV take is windowed separately (at most 30 CSVs per session), so no window crosses a take boundary; 1390 takes shorter than $T=16$ are excluded. Windows use $T=16$, hop 8. The DataPort prose numbers the six activities from 0 to 5, with 4=drink and 5=sit-TV. The downloaded dump’s activity folders skip `action4` and instead use `action5_drink` and `action6_sit_TV`, so folder tokens 5/6 are the same last two classes under a folder-index mismatch, not a second drink label. We canonicalize both the documented 0–5 indices and the dump folder tokens onto {static, walk, fall, sleep, drink, sit_tv}, yielding 2175 windows in the corrected population.

c) *SHARP-80 (AR subset)*: We use Meneghello et al.’s 80 MHz Nexmon channel frequency response (CFR) traces (`.mat`, 242 subcarriers $\times$ 4 monitor antennas) [9], [15]. Folder names alone do not encode subject identity. Their Table 2 maps AR-1, AR-2, AR-7, and AR-8 to persons P1, P2, P3, and P4, respectively [9]. We retain only the campaign-letter-a recordings AR-1a, AR-2a, AR-7a, and AR-8a under those P1–P4 labels. Amplitude from antenna 0 is windowed ($T=100$, hop 50), capped at 6000 packets per file and three files per activity tag, then downsampled to 64 channels, yielding 4760 windows. The eight AR activity tags after preprocessing are empty, walk, run, jump, sit, stand, sitstand (sit-to-stand), and h (arm exercises). Leave-one-subject-out uses these person labels (not campaign letters).

B. Within-Dataset Protocols

Features and model families are identical across protocols:

- **Random (optimistic)**: stratified 70/30 train/test split on windows.

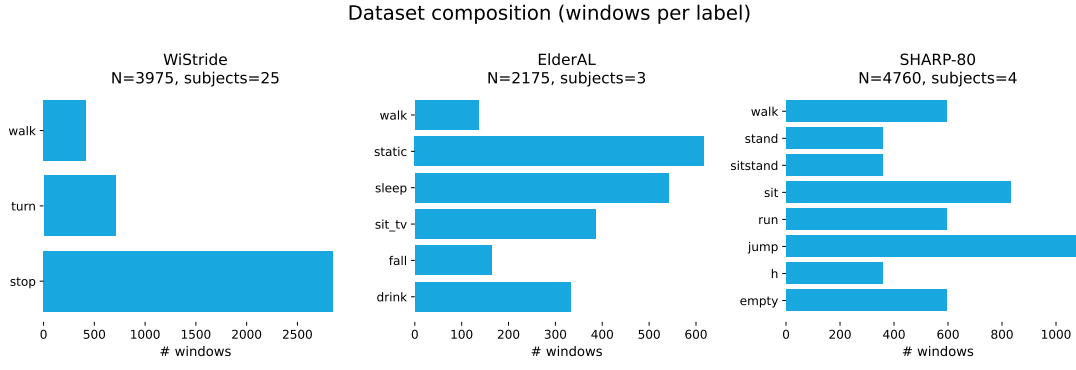


Fig. 1. Dataset composition: number of amplitude windows per activity label after preprocessing. SHARP-80 short tags: empty, walk, run, jump, sit, stand, sitstand (sit-to-stand), and h (arm exercises).

- **Subject-disjoint:** 5-fold **GroupKFold** (grouped by subject) on WiStride and leave-one-subject-out (**LOSO**) evaluation on ElderAL and on SHARP-80 (P1–P4).

For the controlled runs, standard deviations use divisor n ($\text{ddof}=0$) over the declared fits. Neural subject summaries pool three model seeds per fold; this spread combines fold and initialization variation. RF random and repeated-transfer metrics are averaged over split seeds {42, 123, 7, 2026, 31415} and reported as mean $\pm$ standard deviation. Neural metrics use split seeds {42, 123, 7} and model seeds {42, 123, 7}, with each of the 126 fits independently seeded and logged. Subject metrics are reported in the same form over GroupKFold folds (WiStride) or LOSO folds (ElderAL / SHARP-80). The protocol gap

$$\Delta = \text{Acc}_{\text{random}} - \text{Acc}_{\text{subject}} \quad (1)$$

compares the random-window protocol to subject-disjoint evaluation. Because windows overlap ($\text{hop} = T/2$), the random split can also place temporally adjacent windows into both partitions, so Δ is a *combined protocol gap* rather than an isolated measure of subject overlap or temporal leakage. Positive Δ indicates higher accuracy under the random protocol. It does not isolate the cause. Training budgets are reported separately. A training-only matched-budget RF sensitivity analysis is reported in Supplementary Table 1. It downsamples only the training windows, preserves each protocol’s original test membership, and therefore does not remove all protocol-level confounding. The corrected natural-budget counts are recorded per split in the evidence package. The recording-holdout arm holds out complete recordings for ElderAL and SHARP-80. WiStride uses a chronological within-recording split with a raw-sample purge because it has one recording per subject. Every split also reports both a **train-fitted majority classifier** and an **oracle test-label majority reference**. For WiStride RF we additionally report a **subject-cluster bootstrap** 95% confidence interval (CI) on the window-count-weighted protocol gap (subjects resampled as clusters, $B=2000$, seed 42). This descriptive interval is conditional on the saved predictions and does not capture retraining uncertainty. For ElderAL and

SHARP-80 we report complete LOSO fold distributions rather than nominal population CIs over three or four groups.

C. Cross-Dataset Protocol

Cross-dataset results average five repeated source-group selections from **GroupShuffleSplit** seeds {42, 123, 7, 2026, 31415}. For transfer we map every dataset to a shared binary label {walk, non_walk}, which is present on all three corpora (WiStride walk, ElderAL walk, and SHARP-80 w) and avoids the trivial all-motion collapse of a {static, motion} map on WiStride. An RF is trained on each source training partition with $\text{test_size} = 0.2$, which holds out approximately 20% of source subject groups. Because group sizes differ, the resulting fraction of training windows varies by corpus. StandardScaler is fit only on that source training partition. The classifier is evaluated on the *full* target corpus, both zero-shot and after **CORAL** (Correlation Alignment) [17], which matches source and target feature covariances. CORAL estimates target covariance from the full unlabeled target feature matrix before prediction, so the CORAL numbers are *transductive* unsupervised adaptation rather than held-out inductive transfer. We also report an **in-dataset** walk/non-walk subject-safe RF control (Table VII, GroupShuffleSplit 70/30) to check whether the binary task is learnable when train and test share a radio stack. Transfer Acc is always compared to the target majority Acc. We do not claim successful adaptation when transfer Acc does not exceed the majority baseline.

IV. EXPERIMENTAL SETUP

The design idea is simple: treat classifiers as *protocol probes*, not as candidate deployable recognizers. If random-window Acc is inflated for a classical RF and for moderately deep neural baselines under a shared preprocessing pipeline, then random Acc is already an unreliable headline metric for these corpora, regardless of whether a broader benchmark of model families could raise absolute subject-disjoint scores.

A. Features and Models

a) *Random Forest:* Per-subcarrier mean, standard deviation, and maximum over the temporal window yield a 3C-

dimensional vector ($C=64$). We train an RF with 200 trees as a strong non-deep diagnostic of protocol-dependent inflation. The archived **group-ID probe** predicts subject (or SHARP campaign-linked person) labels under contiguous loader-order train/test blocks with up to two window indices omitted on each side of the boundary. This remains a loader-order purge rather than a fully verified physical-time purge, even though the corrected raw-data package now retains recording IDs, sample intervals, and available timestamps for the activity-recognition controls. High group-ID Acc indicates subject/session-correlated structure and serves as a fingerprint diagnostic; it does not isolate biological identity from recording session or campaign, or decompose the headline Δ .

b) CNN1D: A deeper three-block 1D CNN (64 / 128 / 128 channels) with batch normalization, ReLU, max-pooling, adaptive average pooling, and dropout is trained for 25 epochs with Adam (10^{-3} , weight decay 10^{-4}) on preprocessed amplitude windows of shape $T \times 64$, without the temporal summary features used by RF.

c) CNN-GRU: A two-block convolutional front-end followed by a bidirectional GRU and a linear classifier provides a stronger temporal probe under the same training budget.

B. Metrics

We report Acc, macro-F1, and majority-baseline Acc/F1 on every dataset and split. Macro-F1 is especially important on WiStride, where class imbalance already yields majority Acc ≈ 0.72 (Section III).

C. Implementation

RF experiments use scikit-learn. CNN1D and CNN-GRU use PyTorch. The controlled RF package was run locally with the isolated project environment; the 126-fit neural package was run on a Colab Tesla T4. Run IDs are `controlled_raw_20260914` and `neural_raw_20260914`; both packages record environment and source hashes. The full recomputation workflow is in the repository's modular analysis notebooks and CLI mirror. The public Kaggle notebook is an artifact viewer; the current paper tables and figures are tracked in the repository, while full run metadata are retained in the controlled evidence packages and the public companion dataset contains an earlier artifact package. Raw CSI corpora remain on IEEE DataPort. This release does not re-host the original dumps.

V. RESULTS

A. Within-Dataset: Random vs. Subject-Disjoint

Table II reports Acc and macro-F1 with majority Acc. Table III summarizes the protocol gap Δ .

a) WiStride (primary evidence): Given WiStride class imbalance, **macro-F1 is the headline metric**. In the corrected population, RF macro-F1 falls from 0.526 (random) to 0.322 (subject-disjoint GroupKFold), while Acc falls from 0.773 to 0.684 ($\Delta=0.089$). A paired subject-cluster bootstrap on the window-count-weighted RF gap yields a 95% confidence

interval of 6.4–12.1 percentage points (Table III). Subject-disjoint RF Acc sits *below* the majority Acc (≈ 0.72) under that imbalance. The controlled neural probes show the same broad sensitivity: CNN1D falls from Acc 0.663 to 0.396, while CNN-GRU falls from 0.619 to 0.519 across the corresponding seeded fits. The paired 25-subject plot (Fig. 2) shows the per-subject direction of this contrast across the evaluated participants. The combined protocol gap is therefore clear for the RF probe on this 25-subject ESP32 corpus [2], [4].

b) ElderAL (supporting): With only three LOSO folds, ElderAL limits generalization claims. Under the six-activity taxonomy and the corrected no-cross-file population, RF Acc decreases from 0.988 to 0.912 under LOSO ($\Delta=0.076$). Per-held-out Acc is $\{0.900, 0.907, 0.930\}$ (median 0.907). Seeded CNN1D changes from 0.991 to 0.853 ($\Delta=0.139$), while CNN-GRU changes from 0.987 to 0.875 ($\Delta=0.112$). The random-split scores remain high under $T=16$ with hop 8 (50% overlap), but the recording-holdout RF control is also high (0.978), so the corrected results do not attribute the gap to window overlap alone. The quantitative recording-aware and temporal-separation RF controls for all three datasets are reported in Supplementary Table 2. All models remain above the corrected test-majority baseline (≈ 0.285); the small cohort and the exclusion of short takes limit external validity.

c) SHARP-80: Under random splits, RF Acc/F1 reach $0.954 \pm 0.006 / 0.955 \pm 0.006$, then collapse under LOSO to $0.181 \pm 0.039 / 0.122 \pm 0.032$ ($\Delta=0.773$), below the LOSO majority Acc (≈ 0.247). Seeded CNN1D changes from 0.931 to 0.153, and CNN-GRU from 0.896 to 0.140. Near-ceiling Acc under random splits on this Nexmon 80 MHz subset is strongly protocol-dependent for these amplitude baselines. Because each P1–P4 label is campaign-linked, this collapse

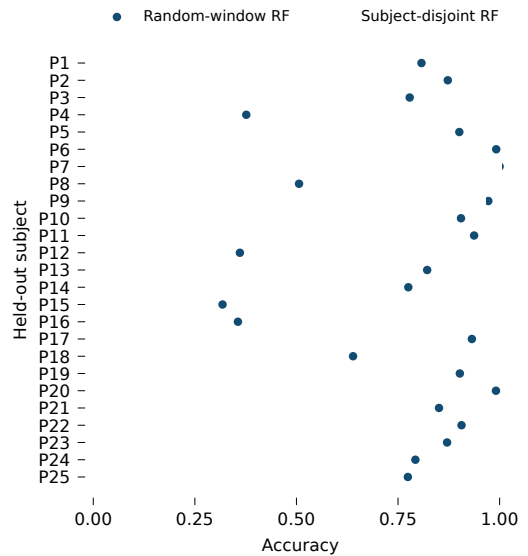


Fig. 2. WiStride RF accuracy for all 25 subjects: random-window versus subject-disjoint evaluation. Lines connect the paired values; no causal identity decomposition is implied.

TABLE II

WITHIN-DATASET ACC AND MACRO-F1 (MEAN $\pm$ STANDARD DEVIATION) FROM THE CORRECTED NO-CROSS-FILE POPULATION. RF RANDOM USES FIVE SPLIT SEEDS; NEURAL ROWS USE THREE SPLIT SEEDS AND THREE INDEPENDENTLY CONTROLLED MODEL SEEDS. WISTRIDE SUBJECT EVALUATION USES FIVE GROUPKFOLD FOLDS. ELDERAL AND SHARP-80 USE LOSO. MAJ. ACC IS THE ORACLE TEST-MAJORITY REFERENCE ON THE SAME TEST FOLDS.

Dataset	Model	Random		Subject		
		Acc	F1	Acc	F1	Subj. Maj*
WiStride	RF	0.773 $\pm$ 0.008	0.526 $\pm$ 0.027	0.684 $\pm$ 0.074	0.322 $\pm$ 0.044	0.716
	CNN1D	0.663 $\pm$ 0.156	0.444 $\pm$ 0.124	0.396 $\pm$ 0.160	0.256 $\pm$ 0.046	0.716
	CNN-GRU	0.619 $\pm$ 0.136	0.372 $\pm$ 0.096	0.519 $\pm$ 0.137	0.298 $\pm$ 0.056	0.716
	Majority	0.717	0.278	0.716 $\pm$ 0.091	0.277 $\pm$ 0.021	—
ElderAL [†]	RF	0.988 $\pm$ 0.003	0.991 $\pm$ 0.003	0.912 $\pm$ 0.013	0.925 $\pm$ 0.011	0.285
	CNN1D	0.991 $\pm$ 0.015	0.993 $\pm$ 0.012	0.853 $\pm$ 0.114	0.869 $\pm$ 0.108	0.285
	CNN-GRU	0.987 $\pm$ 0.021	0.987 $\pm$ 0.023	0.875 $\pm$ 0.066	0.895 $\pm$ 0.053	0.285
	Majority	0.283	0.074	0.285 $\pm$ 0.016	0.074 $\pm$ 0.003	—
SHARP-80 [‡]	RF	0.954 $\pm$ 0.006	0.955 $\pm$ 0.006	0.181 $\pm$ 0.039	0.122 $\pm$ 0.032	0.247
	CNN1D	0.931 $\pm$ 0.017	0.937 $\pm$ 0.015	0.153 $\pm$ 0.063	0.114 $\pm$ 0.054	0.247
	CNN-GRU	0.896 $\pm$ 0.037	0.908 $\pm$ 0.029	0.140 $\pm$ 0.064	0.104 $\pm$ 0.058	0.247
	Majority	0.225	0.046	0.247 $\pm$ 0.025	0.050 $\pm$ 0.004	—
*Test-label majority Acc (mean over subject folds).						
[†] Leave-one-subject-out (3 folds).						
[‡] Leave-one-subject-out (4 subjects, P1–P4), RF random: 5 seeds; neural random: 3 split seeds.						

TABLE III

PROTOCOL GAP $\Delta = \text{Acc}_{\text{random}} - \text{Acc}_{\text{subject}}$ (MEAN). WISTRIDE RF ALSO REPORTS A SUBJECT-CLUSTER BOOTSTRAP 95% CI ON THE WINDOW-COUNT-WEIGHTED GAP.

Dataset	Model	Δ Acc
WiStride	RF	+0.089 [0.064, 0.121]*
	CNN1D	+0.267
	CNN-GRU	+0.100
ElderAL	RF	+0.076
	CNN1D	+0.139
	CNN-GRU	+0.112
SHARP-80	RF	+0.773
	CNN1D	+0.778
	CNN-GRU	+0.756
*Paired subject-cluster bootstrap 95% CI (25 subjects, $B=2000$, seed 42).		

TABLE IV

SHARP-80 RF LEAVE-ONE-SUBJECT-OUT ACC BY HELD-OUT PERSON (P1–P4).

Held-out	RF Acc
P1	0.192
P2	0.229
P3	0.121
P4	0.183
Mean $\pm$ std	0.181 $\pm$ 0.039

is consistent with subject/session shift and should not be read as an isolated biological-identity effect. Table IV reports per-held-out-subject RF Acc (range 0.121–0.229).

d) *Group-ID probe*: Table V reports RF Acc when predicting subject/campaign group IDs from the same amplitude features under purged loader-order blocks. WiStride (25 labels) is the headline setting: RF Acc 0.338 $\pm$ 0.006 far exceeds the majority share (0.040), showing subject/session-correlated structure in separated loader-order blocks. ElderAL and SHARP-80 scores are reported for completeness. With only three and four group labels (and campaign linkage on SHARP), they do not support broad identity-generalization

TABLE V

RF GROUP-ID ACC (MEAN $\pm$ STANDARD DEVIATION OVER SEEDS) UNDER PURGED LOADER-ORDER BLOCKS. MAJ. ACC IS THE LARGEST GROUP'S SHARE.

Dataset	# IDs	RF Acc	Maj Acc
WiStride	25	0.338 $\pm$ 0.006	0.040
ElderAL	3	0.371 $\pm$ 0.013	0.343
SHARP-80	4	0.848 $\pm$ 0.005	0.275

claims or a decomposition of the activity-recognition gap. SHARP group-ID Acc remains high (0.848 $\pm$ 0.005).

B. Empirical Takeaways

- 1) Subject-disjoint splits reduce RF Acc and F1 on the primary WiStride evidence: $\Delta \approx 8.9$ pp (GroupKFold, subject-cluster 95% CI 6.4–12.1 pp) with F1 falling from 0.526 to 0.322. Independently seeded CNN1D and CNN-GRU probes show the same direction of sensitivity, but are model probes rather than deployable recognizers. Support: ElderAL RF $\Delta \approx 7.6$ pp (3 subjects only).
- 2) SHARP-80 LOSO collapses all three probes (0.95 $\rightarrow$ $\approx$ 0.18 for RF), with RF Acc from 0.121 to 0.229 across held-out campaign-linked groups. Labels are campaign-linked.
- 3) The historical population supports group-ID prediction well above majority on WiStride (0.338 vs. 0.040) in a loader-order fingerprint diagnostic.
- 4) Under WiStride imbalance, Acc can fall below majority while macro-F1 still reveals protocol effects.

C. Cross-Dataset Transfer (Walk vs. Non-Walk)

Table VI reports RF zero-shot and CORAL Acc against the target majority Acc. These transfer numbers average five repeated GroupShuffleSplit source seeds; repeated seeds reuse

TABLE VI

CROSS-DATASET RF TRANSFER (WALK VS. NON-WALK) VS. TARGET MAJORITY ACC. ENTRIES ARE MEAN $\pm$ STANDARD DEVIATION OVER FIVE SOURCE GROUPSHUFFLESPLIT SEEDS. CORAL USES FULL-TARGET COVARIANCE AND IS TRANSDUCTIVE. NO PAIR EXCEEDS MAJORITY ACCURACY.

Train $\rightarrow$ Test	RF Acc	CORAL Acc	Maj Acc
WiStride $\rightarrow$ ElderAL	0.063 $\pm$ 0.000	0.924 $\pm$ 0.004	0.937
WiStride $\rightarrow$ SHARP-80	0.145 $\pm$ 0.001	0.864 $\pm$ 0.004	0.875
ElderAL $\rightarrow$ WiStride	0.105 $\pm$ 0.000	0.888 $\pm$ 0.000	0.895
ElderAL $\rightarrow$ SHARP-80	0.148 $\pm$ 0.014	0.860 $\pm$ 0.001	0.875
SHARP-80 $\rightarrow$ WiStride	0.895 $\pm$ 0.000	0.893 $\pm$ 0.001	0.895
SHARP-80 $\rightarrow$ ElderAL	0.937 $\pm$ 0.000	0.934 $\pm$ 0.003	0.937

TABLE VII

IN-DATASET WALK/NON-WALK RF CONTROL. ENTRIES ARE MEAN $\pm$ STANDARD DEVIATION OVER FIVE SUBJECT-SAFE GROUPSHUFFLESPLIT SEEDS; THE MAJORITY REFERENCE IS COMPUTED ON EACH TEST SPLIT.

Dataset	Acc	Maj Acc
WiStride	0.871 $\pm$ 0.023	0.885
ElderAL	0.996 $\pm$ 0.003	0.930
SHARP-80	0.860 $\pm$ 0.020	0.860

the same datasets and are not additional subjects. Table VII is the in-dataset positive control.

No zero-shot pair exceeds the target majority Acc. Two directions match it because they predict the dominant target class by construction. CORAL generally moves the predictions toward the target majority but still does not exceed it: for example, WiStride $\rightarrow$ ElderAL rises from 0.063 to 0.924 against majority 0.937, while WiStride $\rightarrow$ SHARP-80 rises from 0.145 to 0.864 against 0.875. The five-seed spread is small for several majority-like directions but larger for source selections that change the learned minority behavior. CORAL mean macro-F1 slightly exceeds the oracle target-majority reference in all six directions (improvements of approximately 0.0004–0.0076), despite remaining below it in accuracy. The in-dataset control (Table VII) exceeds majority accuracy on ElderAL, falls below it on WiStride, and is effectively tied on SHARP-80. These mixed controls limit how specifically the transfer deficit can be attributed to radio differences.

Across the five source seeds, mean RF zero-shot and CORAL transfer accuracy does not exceed the target-majority reference. The SHARP-80 in-dataset mean exceeds its reference by only 0.00017, which does not establish a reliable advantage.

VI. DISCUSSION

A. Alignment with Prior Protocol Critiques

Our WiStride RF Δ (≈ 8.9 pp under GroupKFold) is consistent in *direction* with person-independent reevaluations of CSI-HAR pipelines [1], [2], [4], measured here on a newer ESP32 gait corpus with 25 subjects. Independently seeded neural probes provide a controlled sensitivity check under the same frozen partitions. SHARP-80 shows a much larger LOSO collapse ($\Delta \approx 77.3$ pp RF Acc). We audit recent multi-radio

DataPort releases under a shared processing pipeline with majority-aware metrics and a group-ID fingerprint diagnostic.

B. Why Acc Alone Can Mislead

On WiStride, subject-disjoint Acc for RF/CNN probes can fall *below* the majority Acc while macro-F1 remains informative under class imbalance. Limited capacity and imbalance can also mask neural gains under random splits (CNN-GRU Acc variance is large). Papers that report only Acc under random splits may therefore understate evaluation risk. We recommend always pairing deep results with a simple subject-disjoint baseline, macro-F1, and a majority reference.

C. Cross-Dataset Reporting

Shared labels must exist on every dataset involved, and transfer Acc must be compared to majority Acc. Matching the majority baseline does not constitute success. Exceeding it is the minimum bar. CORAL can move Acc toward the majority baseline without clearing it, another reason to report the baseline explicitly. In-dataset walk/non-walk controls help separate an ill-posed label map from a hard cross-radio shift.

D. Limitations

- SHARP-80 uses only four AR subjects (P1–P4), and each person label is campaign-linked, so LOSO conflates person and campaign shift.
- ElderAL has only three subjects, which limits generalization claims from its LOSO folds.
- Models are protocol probes (RF, CNN1D, CNN-GRU), not SenseFi-scale architectures. The neural runs use fixed per-fit seeds and 25 epochs, but repeated seeds are not additional subjects and do not provide population-level uncertainty.
- Window lengths differ by dataset (100 vs. 16), reflecting capture length constraints. Overlapping windows confound subject overlap with temporal duplication under random splits; natural training budgets differ, although a training-only matched-budget arm is reported in the controlled RF package. ElderAL takes are kept separate, and its 1390 short takes are excluded; WiStride’s temporal arm uses raw-sample indices and a purge, not a universal physical-time guarantee.
- The group-ID probe uses purged loader-order blocks and cannot separate biological identity from session, campaign, or environment covariates; it is a predictability diagnostic, not a causal decomposition.
- CORAL transfer is transductive (target covariance from the full target set) and reported over five repeated source-group seeds; these seeds reuse the same targets and do not replace independent replication across datasets.
- We use amplitude-only features without phase sanitization or domain-invariant BVP representations as in Widar3.0 [11].

E. Reporting Checklist and Implications

From this multi-radio audit, and consistent with prior subject-partition critiques [1], [2], [4], we recommend:

- 1) Default to subject-disjoint or LOSO evaluation; treat random Acc as an optimistic diagnostic only.
- 2) Disclose the partition rule (subject/window allocation, seeds, folds), window hop, and sampling caps so leakage risk and protocol gaps are auditable.
- 3) Report a majority baseline on the same test folds, and under imbalance lead with macro-F1 (or balanced Acc) rather than Acc alone.
- 4) For cross-dataset claims, use a label map present on every corpus, require Acc to exceed majority on the target, and state whether adaptation is inductive or transductive.
- 5) Dataset releases should ship subject IDs and recommended subject-aware split recipes, not Acc-only leaderboards; pair deep models with a simple subject-disjoint baseline (e.g., RF).

Random-split Acc without these controls is weak evidence of out-of-subject deployability. DataPort-style releases that omit subject IDs or subject-aware split recipes make the optimistic protocol the path of least resistance for downstream papers.

VII. CONCLUSION

Across the three corpora audited here, random-window RF Acc exceeded subject-disjoint Acc, with gaps ranging from 7.6 to 77.3 percentage points in the corrected population. On WiStride, RF Acc falls by ≈ 8.9 pp (subject-cluster 95% CI 6.4–12.1 pp) and macro-F1 drops from 0.526 to 0.322. SHARP-80 LOSO collapses RF Acc from 0.954 to 0.181 under campaign-linked person labels. ElderAL shows a smaller directional gap, but its corrected population excludes short takes and has only three subjects. The three- and four-group results are preliminary, and the protocol gaps do not isolate identity effects. The same windows yield group-ID Acc far above majority on WiStride, and five-seed walk/non-walk transfer across radios does not exceed target majority accuracy even with CORAL. We recommend subject-disjoint evaluation, recording-aware temporal separation, majority baselines, and imbalance-aware metrics as defaults when releasing or benchmarking CSI models. The checklist in Section VI-E summarizes the recommended reporting practices. Future work should evaluate stronger domain-adaptation baselines under the *same* subject-aware protocol.

ACKNOWLEDGMENT

The authors used AI-assisted tools for language editing, formatting suggestions, and assistance writing notebook and analysis code. The research design, experimental protocol, analysis choices, and scientific claims are the authors' own, and the authors take full responsibility for the content.

REFERENCES

- [1] D. Varga, "Critical analysis of data leakage in WiFi CSI-based human action recognition using CNNs," *Sensors*, vol. 24, no. 10, p. 3159, 2024.
- [2] —, "Mitigating data leakage in a WiFi CSI benchmark for human action recognition," *Sensors*, vol. 24, no. 24, p. 8201, 2024.
- [3] —, "Exposing data leakage in Wi-Fi CSI-based human action recognition: A critical analysis," *Inventions*, vol. 9, no. 4, p. 90, 2024.
- [4] D. Varga and A. Q. Cao, "Why partitioning matters: Revealing overestimated performance in WiFi-CSI-based human action recognition," *Signals*, vol. 6, no. 4, p. 59, 2025.
- [5] J. Yang, X. Chen, H. Zou, C. X. Lu, D. Wang, S. Sun, and L. Xie, "SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing," *Patterns*, vol. 4, no. 3, p. 100703, 2023.
- [6] F. Wang, T. Zhang, W. Xi, H. Ding, G. Wang, D. Zhang, Y. Cui, F. Liu, J. Han, J. Xu, and T. X. Han, "A survey on Wi-Fi sensing generalizability: Taxonomy, techniques, datasets, and future research prospects," *IEEE Communications Surveys & Tutorials*, vol. 28, pp. 5227–5266, 2026.
- [7] O. Custance, S. Parkinson, and S. Khan, "WiStride: CSI-based gait dataset from indoor walking subjects," *IEEE DataPort*, 2025, doi: 10.21227/5h8n-wb48.
- [8] J. Li, Z. Huang, X. Cao, C. Song, and S. Chu, "ElderAL-CSI," *IEEE DataPort*, 2025, doi: 10.21227/656t-a531.
- [9] F. Meneghello, N. Dal Fabbro, D. Garlisi, I. Tinnirello, and M. Rossi, "A CSI dataset for wireless human sensing on 80 MHz Wi-Fi channels," *IEEE Commun. Mag.*, vol. 61, no. 9, pp. 146–152, 2023.
- [10] Z. Wang, F. Li, H. Zhao, Z. Mao, Y. Zhang, Q. Huang, B. Cao, M. Cao, B. He, and Q. Hou, "Wi-Fi sensing tool release: Gathering 802.11ax channel state information from a commercial Wi-Fi access point," arXiv:2506.16957, 2025.
- [11] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proc. 17th Annu. Int. Conf. Mobile Systems, Applications, and Services (MobiSys)*. ACM, 2019, pp. 313–325.
- [12] K. Jakkala, A. Bhuyan, Z. Sun, P. Wang, and Z. Cheng, "Deep CSI learning for gait biometric sensing and recognition," arXiv:1902.02300, 2019.
- [13] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, p. 53, 2011.
- [14] J. Schäfer, B. R. Barrsiwal, M. Kokhkharaova, H. Adil, and J. Liebehenschel, "Human activity recognition using CSI information with Nexmon," *Appl. Sci.*, vol. 11, no. 19, p. 8860, 2021.
- [15] F. Meneghello, N. Dal Fabbro, D. Garlisi, I. Tinnirello, and M. Rossi, "CSI dataset for wireless human sensing on 80 MHz Wi-Fi channels," *IEEE DataPort*, 2022, doi: 10.21227/xbhv-f125.
- [16] F. Meneghello, D. Garlisi, N. Dal Fabbro, I. Tinnirello, and M. Rossi, "SHARP: Environment and person independent activity recognition with commodity IEEE 802.11 access points," *IEEE Trans. Mobile Comput.*, vol. 22, no. 10, pp. 6160–6175, 2023.
- [17] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016, pp. 2058–2065.