

# Stability-Screened Source-Pipeline Selection for Later Transfer Validation in Wi-Fi Sensing

Volodymyr Pavlenko  
Lviv Polytechnic National University  
Lviv, Ukraine  
volodymyr.v.pavlenko@lpnu.ua

**Abstract**—Wi-Fi sensing experiments often leave the researcher with several promising source-side pipelines, while time and measurement budget allow later validation of only one or two of them. In this paper, a pipeline denotes a concrete processing chain: for example, CSI packets may be segmented into short windows, simple statistical descriptors may be computed, and the resulting feature vectors may be submitted to a classifier. Another pipeline may use longer windows or a different descriptor set. Selecting the next candidate only by the highest source accuracy is risky because the apparent winner may depend on a favorable split, unstable sessions, or degraded data quality. This paper presents a stability-screened rule for source-pipeline selection before later transfer validation. Each candidate is evaluated by its source-side task score together with three penalties for temporal variation, split sensitivity, and data-quality degradation. The scientific novelty is a source-only screening protocol that treats candidate selection itself as a reproducibility-controlled decision stage, combines these penalties with a frozen no-retuning transfer check, and keeps target data out of the selection step; it does not propose a new sensing model or claim broad transfer robustness. Only the retained candidate is then checked under limited target shifts such as a closed-door setting or a later-day repeat, without target-specific retuning. This helps reduce wasted follow-up experiments, improves reproducibility in early comparison stages, and is applicable to both single-link and cooperative Wi-Fi sensing pipelines.

**Index Terms**—Wireless LAN (IEEE 802.11); Channel state information; Feature extraction; Machine learning; Reliability; Verification and validation.

## I. INTRODUCTION

Wi-Fi sensing studies usually try several candidate pipelines before moving to harder validation conditions. In practice, those candidates may differ in window length, preprocessing, descriptors, or classifier choice, while all of them are built from the same early source data. After this first round, a few candidates often look similarly strong. The problem is that only one or two of them can usually be tested later in changed conditions such as a different room, a closed door, or a new measurement day.

This creates a practical bottleneck. If the next-stage candidate is chosen only by the highest source accuracy, experimental effort can be spent on a pipeline that appears strong for reasons that are not robust. A strong source number may come from a favorable split, unstable sessions, or hidden data-quality issues. For Wi-Fi sensing studies, this is also a reproducibility problem. Recent surveys and systematic studies show that reported performance is strongly shaped by acquisition dis-

cipline, session shifts, and deployment conditions [1]–[5]. Recent protocol-focused studies on Wi-Fi CSI evaluation also show that improper partitioning can introduce strong leakage effects and inflate reported performance [6], [7]. More general work on evaluation practice also shows that unstable partitions can make a weak candidate look stronger than it really is [8], [9]. In the broader out-of-distribution literature, evaluation-protocol studies likewise caution against oracle-style selection and hidden test-domain information leakage [10]. This concern is also consistent with the recent ESP32-S3 CSI/RSSI indoor ranging study, which treated the acquisition pipeline itself as an object of experimental evaluation and showed that low-cost Wi-Fi measurements require careful protocol design and validation [4].

This paper addresses a narrower question than general transfer robustness: when several source-side candidates look acceptable, how should one candidate be chosen for the next validation step? The answer proposed here is intentionally small and practical. The scientific novelty is not a new sensing algorithm or a target-adaptation scheme. Instead, the paper contributes an explicit pre-transfer decision layer for Wi-Fi sensing workflows: a source-only score that separates temporal instability, split sensitivity, and data-quality degradation, together with a frozen no-retuning validation path for the retained candidate. Relative to accuracy-only ranking or a generic dispersion-only heuristic, this formulation keeps the decision interpretable and tied to concrete Wi-Fi acquisition risks.

## II. PROBLEM FORMULATION

In this paper, a candidate denotes a complete processing pipeline that takes Wi-Fi measurements and produces a final decision. For example, a candidate may take CSI packets, form 1-second windows, compute mean and variance features, and classify them into two classes such as motion/no-motion. Another candidate may use 2-second windows, a different descriptor set, or another classifier.

Assume that an initial Wi-Fi sensing study has already produced several reasonable candidates of this kind. One candidate may use short CSI windows and simple statistics, another may use longer windows, and a third may use a different descriptor set. In all three cases the processing chain is explicit: Wi-Fi packets are collected, windows are formed, features are computed, and a classifier produces the final decision. All of

them can look acceptable on source data. However, checking every candidate under room changes and later sessions is costly in time, labeling, and result interpretation.

The real risk is not only lower transfer performance later. The deeper risk is making an unjustified choice too early. A candidate may appear best only because its source result depends on unstable windows, favorable partitioning, or silent quality problems in the data. If that candidate is selected, a later failure can be attributed to the room change even though the weaker point was already visible in the source stage.

The problem studied here is therefore simple: from several acceptable source-side candidates, choose one candidate that is credible enough to take into the next validation step. The goal is not to maximize a single source score at any cost. The goal is to make a better practical choice before additional experiments are spent.

### III. PROPOSED METHOD

The proposed approach has two stages. First, all candidate pipelines are screened on source data. Second, only the retained candidate is frozen and checked under a small number of target shifts.

The source-side screening uses grouped splits. Operationally, a whole acquisition session stays on one side of the split. If sliding windows are used, overlapping raw-packet intervals are not allowed to cross the train/test boundary. This reduces leakage and makes the source comparison statistically cleaner.

The grouped-split protocol follows four ordered steps. First, define the group key before any split is made; by default the group is the acquisition session, and, if subject identifiers are available, the recommended compound key is (subject, session). Second, assign whole groups to source folds. Third, generate windows only after the split, separately inside each fold. Fourth, compute features, train the classifier, and evaluate on the resulting fold-specific windows. This order prevents packet overlap or session fragments from crossing the split boundary.

Each candidate  $c$  receives a combined source-side score

$$J(c) = A_{\text{src}}(c) - \lambda_t D_{\text{temp}}(c) - \lambda_s D_{\text{split}}(c) - \lambda_q D_{\text{qual}}(c), \quad (1)$$

where  $A_{\text{src}}(c)$  is the main source metric, such as balanced accuracy or macro F1. The three penalty terms have distinct interpretations:  $D_{\text{temp}}(c)$  penalizes candidates whose source performance changes too much over time,  $D_{\text{split}}(c)$  penalizes candidates that depend too strongly on how the train/test split was chosen, and  $D_{\text{qual}}(c)$  penalizes candidates that exhibit degraded data quality, for example an elevated invalid-window rate or timestamp-repair rate. The weights  $\lambda_t$ ,  $\lambda_s$ , and  $\lambda_q$  define how strongly these instability indicators should reduce the final score.

Operationally, the rule is straightforward. Start from the source-side score. Then subtract a penalty if the candidate is unstable across repeated source sessions. Subtract another penalty if the result changes too much when the split changes. Subtract a third penalty if the data associated with that

candidate show signs of degradation. After that, compare the final scores and retain the candidate with the best combined value.

Let  $a_k(c)$  denote the source metric on grouped split  $k$ , and let  $A_{\text{src}}(c) = K^{-1} \sum_{k=1}^K a_k(c)$  be the corresponding source mean. Define the normalized penalty terms as

$$D_{\text{split}}(c) = \text{clip}\left(\frac{\text{std}_k(a_k(c))}{\tau_s}, 0, 1\right), \quad (2)$$

$$D_{\text{temp}}(c) = \text{clip}\left(\frac{|A_{\text{early}}(c) - A_{\text{late}}(c)|}{\tau_t}, 0, 1\right), \quad (3)$$

$$D_{\text{qual}}(c) = \text{clip}(w_{\text{inv}}r_{\text{inv}}(c) + w_{\text{mis}}r_{\text{mis}}(c) + w_{\text{rep}}r_{\text{rep}}(c), 0, 1), \quad (4)$$

where  $A_{\text{early}}(c)$  and  $A_{\text{late}}(c)$  are the mean source metrics on early and late source-session groups,  $r_{\text{inv}}(c)$  is the invalid-window rate,  $r_{\text{mis}}(c)$  is packet-missingness growth measured relative to the nominal packet count for the same windowing recipe, and  $r_{\text{rep}}(c)$  is the timestamp-repair rate. The quality weights satisfy  $w_{\text{inv}} + w_{\text{mis}} + w_{\text{rep}} = 1$ .

The screening parameters are also fixed by a deterministic source-only rule. Unless there is a predeclared domain-specific priority, use  $\lambda_t = \lambda_s = \lambda_q = \frac{1}{3}$ . Let  $\tau_s$  be the median split-dispersion value across all candidates in the source study and let  $\tau_t$  be the median early/late gap across the same candidate set. For the stability check, evaluate the weight grid  $\{0.8, 1.0, 1.2\}\lambda_t \times \{0.8, 1.0, 1.2\}\lambda_s \times \{0.8, 1.0, 1.2\}\lambda_q$ , renormalize each triplet to sum to 1, and mark the selection unstable if the winner changes in more than 25% of the grid points. These quantities are fixed before any target result is inspected.

The retained candidate is  $c^* = \arg \max_c J(c)$ . This rule is still simple, but it differs from two baseline choices. Accuracy-only ranking keeps only  $A_{\text{src}}$  and ignores instability indicators. A generic mean-variance heuristic can penalize overall dispersion, but it does not separate temporal drift, split sensitivity, and data-quality degradation. The novelty here is the explicit decomposition of these three penalties under fixed source-only weights, followed by a frozen no-retuning target check for only the retained candidate.

A simple example illustrates the intended use. Suppose candidate A uses short CSI windows and achieves the highest source accuracy, but its result changes substantially between repeated sessions and between different train/test splits. Candidate B uses slightly longer windows, starts from a slightly lower source accuracy, but is much more stable and has cleaner data. The proposed rule is designed exactly for that case: it allows B to be retained if A appears strong mainly because of instability or favorable partitioning.

Table I is an illustrative source-side example, not a measured transfer result. It is included only to show how the rule behaves. The table can be read directly. Candidate A starts from 0.89, but after subtracting larger penalties its final score becomes 0.79. Candidate B starts lower, at 0.86, but loses much less to penalties and finishes at 0.84. Candidate C is stable as well, but its starting score is lower, so it remains

TABLE I. ILLUSTRATIVE SOURCE-SIDE RANKING UNDER FIXED SOURCE-ONLY WEIGHTS

Cand.	$A_{src}$	$D_t$	$D_s$	$D_q$	$J$
A	0.89	0.34	0.28	0.12	0.79
B	0.86	0.08	0.05	0.04	0.84
C	0.84	0.10	0.07	0.03	0.81

below B. For comparison, the baseline policy selects the candidate with the highest source-side performance and does not apply screening penalties. Figure 1 shows the proposed screened-selection workflow only: source-side grouped-split evaluation, penalty computation, screened-score construction, candidate retention, and frozen validation.

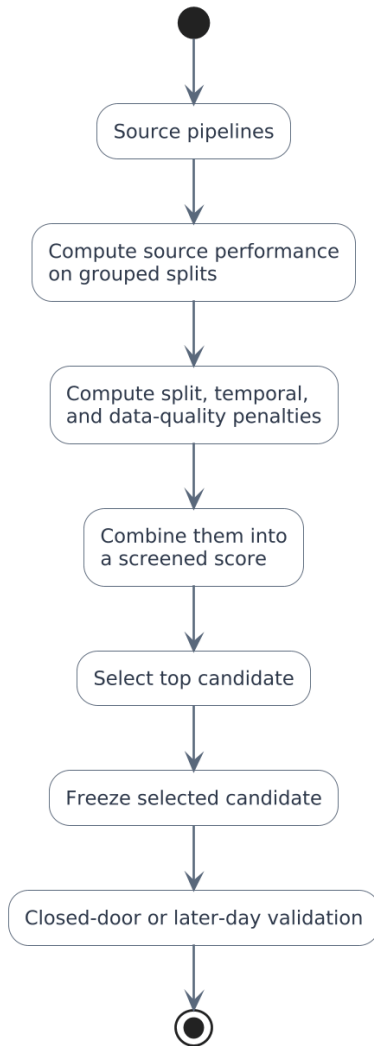


Fig. 1. Workflow diagram for the proposed screened-selection pipeline before the frozen closed-door or later-day validation step.

#### IV. VALIDATION PATH

The post-selection validation stage is deliberately limited. The retained candidate  $c^*$  is frozen and checked under two

target conditions that are realistic yet still compact enough for the present paper. Frozen elements include signal choice, preprocessing, descriptor set, window size, classifier family, hyperparameters, decision thresholds, fusion logic, and the declared  $\lambda$ -weights. Allowed target-side operations are limited to packet parsing, timestamp repair, and predeclared missing-window filters that do not inspect target labels. The target checks are:

- a two-room closed-door setting, in which the propagation path changes through a closed doorway or wall while the task labels, node roles, and acquisition recipe are held fixed;
- a later-day or controlled re-placement repeat, in which the acquisition recipe is unchanged while only the acquisition day or node placement is changed in a controlled way.

The main hypothesis is not that the retained candidate will remain universally strong. The more precise claim is smaller: a stability-screened rule should choose a more defensible next-step candidate than naïve source-accuracy ranking. In the smallest useful validation,  $c^*$  should be compared only against the accuracy-only source winner. Prior work on robust motion-tracking and fused RSSI/RTT localization shows that transfer gains in shifted environments are mostly achieved through explicit robustness controls and quality-aware candidate management [11], [12]. Practical value would be indicated if the stability-screened candidate exhibits smaller source-to-target degradation or more interpretable failure behavior under the same target checks. If both candidates fail similarly, that outcome is still informative because it shows that the current source-screening variables are not sufficient and that the next research step should revise the descriptors or acquisition policy.

#### V. RESEARCH VALUE AND LIMITS

The proposed topic is intentionally narrow. It does not claim a new Wi-Fi sensing modality, a universal transfer benchmark, or a finished cross-environment result package. Its value is simpler: it adds a small and useful decision layer between early source experiments and later transfer validation. In many Wi-Fi sensing studies, this is exactly where effort is lost. Several candidates look good, but there is no clear rule for deciding which one deserves the next expensive check.

The approach is also easy to reuse. The same screening rule can be applied to single-link baselines, local-feature pipelines, or cooperative pipelines, provided that the measurement protocol and quality penalties are reported clearly. A later publication can study whether the retained candidate truly generalizes better. The present paper makes a smaller claim: before that larger study begins, the selection step itself should be more disciplined and easier to explain. In short, the paper is about making a better practical choice at the moment when several pipelines still look plausible.

*Threats to validity:* Three threats should be stated explicitly. First, screening on source data may still induce selection bias; this is mitigated here by grouped splits, explicit

dispersion penalties, and mandatory reporting of the source-side scoring rule. Second, the penalties are heuristic indicators rather than guarantees of target performance; their practical value must therefore be checked empirically and is not assumed a priori. Third, the outcome can depend on the group definition (session only versus subject-session grouping), so the chosen group key should be reported explicitly and sensitivity to an alternative grouping should be checked when the data allow it.

## VI. CONCLUSION

This work reframes a small but recurring Wi-Fi sensing difficulty as a selection problem: before transfer validation begins, the research process needs a principled way to decide which source-conditioned candidate is worth carrying forward. The proposed stability-screened rule offers that intermediate step by combining source task quality with explicit penalties for instability, split sensitivity, and data-quality degradation in one score. The rule reduces wasted validation effort, strengthens reproducibility in controlled Wi-Fi measurement studies, and improves the interpretability of later transfer results.

## ACKNOWLEDGMENT

The paper draws on materials and intermediate results from the project "Intelligent Methods and Tools for Designing Modules for Autonomous Cyber-Physical Systems" (state registration No. 0124U002340, 2024–2028, Lviv Polytechnic National University).

## REFERENCES

- [1] M. Cominelli, F. Gringoli, and F. Restuccia, "Exposing the CSI: A systematic investigation of csi-based wi-fi sensing capabilities and limitations," in *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2023, pp. 81–90. [Online]. Available: <https://dblp.org/rec/conf/percom/CominelliGR23.html>
- [2] C. Chen, G. Zhou, and Y. Lin, "Cross-domain wifi sensing with channel state information: A survey," *ACM Computing Surveys*, vol. 55, no. 11, pp. 231:1–231:37, 2023. [Online]. Available: <https://dblp.org/rec/journals/csur/ChenZL23.html>
- [3] Z. Shi, J. A. Zhang, R. Y. D. Xu, and Q. Cheng, "Environment-robust device-free Human Activity Recognition with Channel-State Information enhancement and one-shot learning," *IEEE Transactions on Mobile Computing*, vol. 21, no. 2, pp. 540–554, 2022.
- [4] M. Diadiuk and V. Pavlenko, "Design and experimental evaluation of CSI and RSSI-based indoor Wi-Fi ranging on ESP32-S3," *Information and Communication Technologies, Electronic Engineering*, vol. 6, no. 1, pp. 86–100, 2026. [Online]. Available: <https://doi.org/10.23939/ictee2026.01.086>
- [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. [Online]. Available: <https://doi.org/10.1016/j.patter.2023.100703>
- [6] D. Varga, A. Csordás, J. Sütő, and K. Varga, "Mitigating data leakage in a WiFi CSI benchmark for human action recognition," *Sensors*, vol. 24, no. 24, p. 8201, 2024. [Online]. Available: <https://www.mdpi.com/1424-8220/24/24/8201>
- [7] —, "Critical analysis of data leakage in WiFi CSI-based human action recognition using CNNs," *Sensors*, vol. 24, no. 10, p. 3159, 2024. [Online]. Available: <https://www.mdpi.com/1424-8220/24/10/3159>
- [8] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, p. 91, 2006. [Online]. Available: <https://doi.org/10.1186/1471-2105-7-91>
- [9] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," *Journal of Cheminformatics*, vol. 6, p. 10, 2014. [Online]. Available: <https://doi.org/10.1186/1758-2946-6-10>
- [10] H. Yu, X. Zhang, R. Xu, J. Liu, Y. He, and P. Cui, "Rethinking the evaluation protocol of domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 21 897–21 908. [Online]. Available: <https://doi.org/10.1109/CVPR52733.2024.02068>
- [11] D. Wu, Y. Zeng, R. Gao, S. Li, Y. Li, R. C. Shah, H. Lu, and D. Zhang, "WiTraj: Robust Indoor Motion Tracking with WiFi Signals," *IEEE Transactions on Mobile Computing*, vol. 22, no. 5, pp. 3062–3078, 2023. [Online]. Available: <https://doi.org/10.1109/TMC.2021.3133114>
- [12] H. Rizk, A. M. Elmogy, and H. Yamaguchi, "A Robust and Accurate Indoor Localization using Learning-Based fusion of Wi-Fi RTT and RSSI," *Sensors*, vol. 22, no. 7, 2022. [Online]. Available: <https://doi.org/10.3390/s22072700>