BENCHMARK-RELATIVE TEMPORAL-RESOLUTION SUFFICIENCY FOR WEARABLE TIME-SERIES LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A wearable system must choose temporal resolution before a downstream representation can exploit the signal. Conventional resolution sweeps rank tested configurations, but they do not define when a cheaper rate–representation pair preserves enough utility to remain acceptable. We formulate this as *benchmark-relative temporal-resolution sufficiency*. A declared tolerance first defines an admissible set relative to the best tested reference; deployment costs are considered only inside that set. This yields a single decision chain from *admissibility* to *preference* to *validity*. We characterize how the admissible set changes with tolerance, reference strength, target environments, and budgets; give an exact finite-grid extraction algorithm; separate oracle information loss from fitted-model reversals; and derive a uniform-error condition under which the empirical sufficient set and frontier are stable. Across UCI HAR, HAR70+, and Daphnet, a one-point macro-F1 tolerance changes the minimum tested rate from best-score references at 10, 10, and 64 Hz to sufficient rates at 5, 10, and 32 Hz. Class-specific, cross-cohort, perturbation, data-volume, and cost analyses then identify when those decisions cease to be adequate. The contribution is therefore not a universal sampling threshold, but a formal account of when a finite resolution comparison supports a lower-cost design decision and what evidence is required for that decision to remain valid.

1 INTRODUCTION

A wearable sensing system commits to an acquisition rate before any downstream representation can recover or exploit temporal detail. Higher rates preserve a richer observation, but they also increase acquisition, storage, communication, and often processing burden. The relevant design question is therefore not simply which tested rate scores highest, but whether the extra utility obtained at a higher rate is large enough to matter for the deployment decision.

A conventional resolution sweep leaves that question underspecified. It produces a ranking, yet a deployment decision additionally needs an admissible loss in utility, a reference against which that loss is measured, the environments in which preservation must hold, and the costs that matter once utility is adequate. These ingredients induce three distinct stages: *score* identifies the best tested configuration, *sufficiency* identifies configurations that remain acceptably close to that reference, and *preference* compares costs only among those sufficient configurations. The same score table can therefore imply different design choices when the tolerance, reference set, target set, or measured costs change (Anguita et al., 2013; Ustad et al., 2023; Liu et al., 2024).

We formalize this structure through benchmark-relative representation sufficiency. A configuration combines a sampling rate and a representation. Relative to a declared comparison set, it is ϵ-sufficient when its utility gap from the best tested reference is at most ϵ; the *sufficiency frontier* is the cost-nondominated subset of those admissible configurations. This ordering—utility first, cost second—is the core object of the paper. Extending the same rule target by target yields a robust admissible set under environment shift.

The theoretical development follows the same decision chain. Section 3 characterizes the admissible set itself: relaxing the tolerance can only enlarge it, strengthening the reference shifts every shared candidate by the same reference increment, adding targets can only make robust sufficiency harder,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

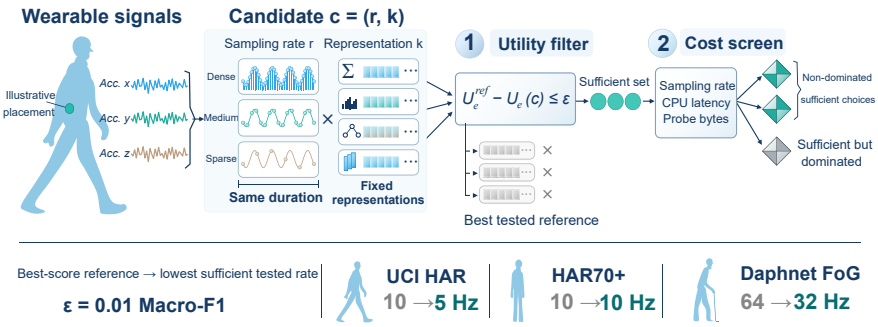


Figure 1: Wearable temporal-resolution selection: compare rate–representation pairs, retain those meeting a declared utility tolerance, and assess their measured costs. The schematic illustrates the decision context rather than a universal sampling threshold.

and a device budget is feasible exactly when its best attainable utility remains within tolerance of the fixed reference. Algorithm 1 implements this finite-grid decision rule exactly. Section 4 then asks when the resulting empirical decision is meaningful. Deterministic coarsening cannot improve unrestricted oracle risk, yet a restricted fitted learner can reverse that ordering when reductions in approximation and fitting error outweigh the oracle information penalty. Finally, a uniform utility-error bound yields a set sandwich and a margin condition for stability of the empirical frontier (**?**Nie et al., 2023).

The experiments instantiate these consequences rather than form a separate collection of ablations. On UCI HAR, HAR70+, and Daphnet, winner selection points to 10, 10, and 64 Hz, while a one-point macro-F1 tolerance yields minimum sufficient tested rates of 5, 10, and 32 Hz. Tolerance and reference changes move that boundary in the directions predicted by the theory; representation choice affects which low-rate configurations remain admissible; and stricter class, target, perturbation, or cost requirements reveal where the aggregate within-cohort decision stops being adequate. These values are finite-grid examples of the decision rule, not universal wearable sampling limits.

2 Related Work

Our problem sits between temporal representation learning, information preservation, and multi-objective evaluation, but differs from each in the object being decided.

Temporal representation versus acquisition resolution. TNC, REBAR, RelCon, and HiMAE learn useful wearable representations from local, relative, or multi-resolution temporal structure (Tonekaboni et al., 2021; Xu et al., 2024a; 2025; Lee et al., 2026). PatchTST, TimesNet, TimeMixer, Pathformer, CoST, iTransformer, and FITS likewise exploit patches, periods, multiple scales, or compact temporal parameterizations (Nie et al., 2023; Wu et al., 2023; Wang et al., 2024; Chen et al., 2024; Woo et al., 2022; Liu et al., 2024; Xu et al., 2024b). These methods alter how a recorded signal is represented. Our intervention is earlier: we alter the sampled input itself while holding physical window duration fixed, and ask which tested rate–representation pairs still preserve declared downstream utility. The frozen random-convolution bank used in our grid is only a fixed probe inspired by ROCKET (Dempster et al., 2020).

Preservation and environment shift. Information-bottleneck methods formalize retention of task-relevant information (Alemi et al., 2017; Federici et al., 2020). Our use of "sufficiency" is deliberately narrower: it is a benchmark-relative predictive-utility condition, not a claim of statistical sufficiency or conditional independence. DomainBed and group DRO motivate evaluating decisions across environments or groups (Gulrajani & Lopez-Paz, 2021; Sagawa et al., 2020); our target-wise extension uses that perspective to ask whether the same admissibility condition survives a declared target rather than proposing a new robust-training algorithm.

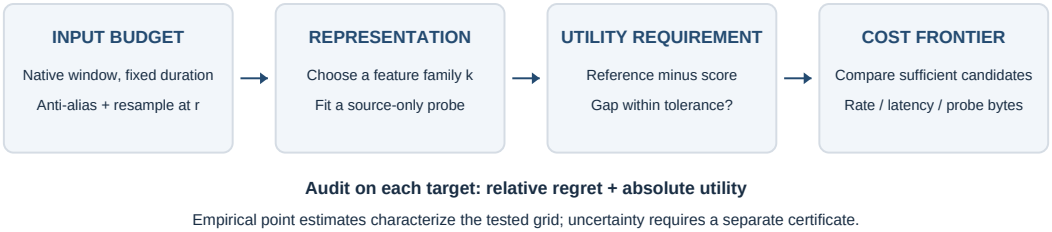


Figure 2: Decision workflow. Input resolution and representation define candidates; the utility requirement determines admissibility; measured costs determine preference within the admissible set.

Frontiers and competing objectives. Pareto analysis is useful when objectives should remain explicit rather than collapsed into a single score. Wang et al. derive a theoretical regret–inference frontier for bandits with network interference (Wang et al., 2025). Their frontier characterizes a fundamental trade-off over policies. Ours is different: it is a finite empirical frontier formed *after* a utility-feasibility filter. The central contribution is therefore not a new Pareto optimizer, but a formal decision layer that converts a resolution sweep into an auditable preservation-and-cost choice.

3 Decision Theory of Benchmark-Relative Sufficiency

The theory starts with the decision itself. A resolution comparison is not fully specified by its scores; it also requires the alternatives being compared, the utility metric, the admissible loss, the costs used for preference, and the environments in which preservation is required. We collect these choices in

$$\Gamma = (\mathcal{C}, \mathsf{M}, \epsilon, \mathbf{b}, \mathcal{E}_{\mathrm{tar}}),$$

where $\mathcal{C}$ is a finite candidate set, M is a higher-is-better utility metric, $\epsilon \geq 0$ is the acceptable utility loss, $\mathbf{b}$ contains the measured cost coordinates, and $\mathcal{E}_{\mathrm{tar}}$ is the declared target set. The section then follows the decision in order: define admissibility, characterize how the contract moves the admissible set, impose cost preference, extend the requirement across targets, and extract the result exactly.

3.1 From scores to an admissible set

Let $e \in \mathcal{E}$ index an environment and $(X, Y) \sim P_e$ denote a native-rate sensor window and its task label. For fixed physical duration T, the deterministic operator S_r produces an anti-aliased view at sampling rate r. A configuration $c = (r, k) \in \mathcal{C}$ applies representation ϕ_k and a fixed fitting procedure to obtain a predictor $\widehat{h}_{s,c}$ from source data D_s. Its population utility on target e is

$$U_{s\to e}(c) = \mathsf{M}\Big(P_e; \widehat{h}_{s,c} \circ \phi_k \circ S_r\Big), \qquad \widehat{U}_{s\to e}(c) = \text{its evaluation estimate.} \tag{1}$$

Within-cohort evaluation abbreviates this notation to $U_e(c)$. Macro-F1 is treated as a functional of the joint label–prediction distribution rather than an additive per-window loss.

For an explicit comparison set $\mathcal{C}$, define

$$U^{\mathrm{ref}}_{e,\mathcal{C}} = \max_{c\in\mathcal{C}} U_e(c), \qquad \Delta_{e,\mathcal{C}}(c) = U^{\mathrm{ref}}_{e,\mathcal{C}} - U_e(c), \qquad \mathcal{S}_{e,\mathcal{C}}(\epsilon) = \{c \in \mathcal{C} : \Delta_{e,\mathcal{C}}(c) \leq \epsilon\}. \tag{2}$$

Definition 1 (Benchmark-relative sufficiency)**.** *A configuration is ϵ-sufficient for $(e, \mathcal{C})$ if it belongs to $\mathcal{S}_{e,\mathcal{C}}(\epsilon)$. Replacing population utility by $\widehat{U}$ defines the empirical set $\widehat{\mathcal{S}}_{e,\mathcal{C}}(\epsilon)$.*

The definition is intentionally operational. It asks whether a tested configuration preserves declared predictive utility relative to a tested reference; it does not assert $Y \perp X \mid \phi_k(S_r(X))$ or equality with Bayes-optimal performance.

3.2 How the contract moves the admissible set

The first result characterizes two changes that recur throughout the paper: relaxing the tolerance and strengthening the reference (Roggen et al., 2010).

Proposition 1 (Tolerance and reference-set order)**.** *Fix environment e and population utilities on a finite candidate set.*

(i) If $0 \le \epsilon_1 \le \epsilon_2$, then $\mathcal{S}_{e,\mathcal{C}}(\epsilon_1) \subseteq \mathcal{S}_{e,\mathcal{C}}(\epsilon_2)$.

(ii) Let $\mathcal{C} \subseteq \mathcal{C}'$ and assume the utilities of all shared configurations are unchanged. Define

$$\delta_e(\mathcal{C}, \mathcal{C}') = U^{\mathrm{ref}}_{e,\mathcal{C}'} - U^{\mathrm{ref}}_{e,\mathcal{C}} \ge 0.$$

Then every shared candidate satisfies

$$\Delta_{e,\mathcal{C}'}(c) = \Delta_{e,\mathcal{C}}(c) + \delta_e(\mathcal{C}, \mathcal{C}'), \qquad c \in \mathcal{C}, \tag{3}$$

and therefore

$$\mathcal{S}_{e,\mathcal{C}'}(\epsilon) \cap \mathcal{C} = \{c \in \mathcal{C} : \Delta_{e,\mathcal{C}}(c) \le \epsilon - \delta_e(\mathcal{C}, \mathcal{C}')\}, \tag{4}$$

with the right-hand side empty when $\epsilon < \delta_e(\mathcal{C}, \mathcal{C}')$.

Thus tolerance and reference strength act in opposite directions. Relaxing ϵ can only admit candidates, while improving the best tested reference consumes the same amount of tolerance for every shared candidate. On a finite grid this immediately structures the rate boundary. Define

$$r_{\min,e}(\epsilon) = \min\{r : (r,k) \in \mathcal{S}_{e,\mathcal{C}}(\epsilon)\}.$$

Then $r_{\min,e}(\epsilon)$ is a nonincreasing step function whose breakpoints lie in the finite set $\{\Delta_{e,\mathcal{C}}(c) : c \in \mathcal{C}\}$. Every jump in a tolerance-sensitivity curve therefore corresponds to a measured candidate gap; no continuous threshold between tested rates is implied.

3.3 From admissibility to cost preference

Utility decides whether a configuration remains in consideration. Only after that filter do deployment costs decide preference. Let $\mathbf{b}(c) \in \mathbb{R}^d_+$ collect measured costs, with smaller coordinates preferred, and write $c' \prec_b c$ when $\mathbf{b}(c') \le \mathbf{b}(c)$ coordinatewise with at least one strict inequality. The sufficient frontier is

$$\mathcal{F}_{e,\mathcal{C}}(\epsilon) = \{c \in \mathcal{S}_{e,\mathcal{C}}(\epsilon) : \nexists c' \in \mathcal{S}_{e,\mathcal{C}}(\epsilon) \text{ with } c' \prec_b c\}. \tag{5}$$

This is a Pareto set over a finite utility-feasible grid, not a fundamental lower bound over all possible representations or policies.

The same decision can be viewed from a budget rather than from the frontier. For a coordinatewise budget $\mathbf{B}$, let

$$\mathcal{C}(\mathbf{B}) = \{c \in \mathcal{C} : \mathbf{b}(c) \le \mathbf{B}\}, \qquad V_e(\mathbf{B}) = \max_{c \in \mathcal{C}(\mathbf{B})} U_e(c),$$

with $V_e(\mathbf{B}) = -\infty$ if the budget set is empty.

Proposition 2 (Budget–frontier equivalence)**.** *For any nonempty budget set $\mathcal{C}(\mathbf{B})$,*

$$\mathcal{C}(\mathbf{B}) \cap \mathcal{S}_{e,\mathcal{C}}(\epsilon) \ne \varnothing \quad \Longleftrightarrow \quad U^{\mathrm{ref}}_{e,\mathcal{C}} - V_e(\mathbf{B}) \le \epsilon. \tag{6}$$

The frontier and budget views therefore encode the same preservation requirement from opposite directions. The reference remains fixed to the original comparison set: recomputing it inside each budget would answer the weaker question of which candidate is best under that budget and would make every nonempty budget contain a zero-gap winner (Davis & Owusu, 2016; Cover & Thomas, 2006).

3.4 Across environments: worst-target sufficiency

A configuration that is adequate in one environment need not remain adequate in another. For a fixed source s and a declared set of compatible targets $\mathcal{E}_{\mathrm{tar}}(s)$, define

$$\Delta_{s\to e,\mathcal{C}}(c) = \max_{c' \in \mathcal{C}} U_{s\to e}(c') - U_{s\to e}(c),$$

and the worst-target gap

$$\Delta^{\mathrm{rob}}_{s,\mathcal{C}}(c) = \max_{e \in \mathcal{E}_{\mathrm{tar}}(s)} \Delta_{s\to e,\mathcal{C}}(c). \tag{7}$$

Algorithm 1 Exact finite-grid sufficiency frontier

Require: finite candidates $\mathcal{C}$; target set $\mathcal{E}_{\mathrm{tar}}$; utilities $\widehat{U}_e(c)$; tolerance ϵ; costs $\mathbf{b}(c)$; optional floors τ_e

Ensure: empirical sufficient set $\widehat{\mathcal{S}}$, nondominated frontier $\widehat{\mathcal{F}}$, and minimum sufficient tested rate $\widehat{r}_{\min}$

for each target $e \in \mathcal{E}_{\mathrm{tar}}$ **do**
  $\widehat{U}_e^{\mathrm{ref}} \leftarrow \max_{c \in \mathcal{C}} \widehat{U}_e(c)$
end for
for each candidate $c \in \mathcal{C}$ **do**
  $g(c) \leftarrow \max_{e \in \mathcal{E}_{\mathrm{tar}}} \{\widehat{U}_e^{\mathrm{ref}} - \widehat{U}_e(c)\}$
  mark c admissible iff $g(c) \le \epsilon$ and all declared floors are met
end for
$\widehat{\mathcal{S}} \leftarrow$ all admissible candidates; $\widehat{\mathcal{F}} \leftarrow \widehat{\mathcal{S}}$
for each $c \in \widehat{\mathcal{S}}$ **do**
  if there exists $c' \in \widehat{\mathcal{S}}$ with $\mathbf{b}(c') \le \mathbf{b}(c)$ coordinatewise and one strict inequality **then**
    remove c from $\widehat{\mathcal{F}}$
  end if
end for
$\widehat{r}_{\min} \leftarrow \min\{r : (r,k) \in \widehat{\mathcal{S}}\}$ if $\widehat{\mathcal{S}} \neq \varnothing$
return $(\widehat{\mathcal{S}}, \widehat{\mathcal{F}}, \widehat{r}_{\min})$

Proposition 3 (Target-wise characterization)**.** *The robust sufficient set is*

$$\mathcal{S}_{\mathrm{rob},s,\mathcal{C}}(\epsilon) = \bigcap_{e \in \mathcal{E}_{\mathrm{tar}}(s)} \mathcal{S}_{s \to e,\mathcal{C}}(\epsilon) = \{c \in \mathcal{C} : \Delta^{\mathrm{rob}}_{s,\mathcal{C}}(c) \le \epsilon\}. \tag{8}$$

Increasing ϵ can only enlarge this set, whereas enlarging the declared target set can only shrink it. If target-specific absolute floors τ_e are required, feasibility additionally requires $U_{s \to e}(c) \ge \tau_e$ for every target.

The max operator exposes the relevant failure mode: average target performance can look acceptable while one target violates the preservation requirement. The robust set can also be empty even when each target has a nonempty sufficient set separately, because the target-wise sets need not share a common configuration.

3.5 Exact finite-grid decision rule

The preceding definitions can be evaluated directly; no iterative optimization is required.

Theorem 1 (Correctness and complexity of Algorithm 1)**.** *Assume complete utility values for all declared candidate–target pairs and complete cost vectors in d measured coordinates. Algorithm 1 returns exactly the empirical set in Equation* (8) *(or Equation* (2) *for a single target), exactly its cost-nondominated subset in Equation* (5)*, and the minimum sufficient tested rate. With $N = |\mathcal{C}|$ and $E = |\mathcal{E}_{\mathrm{tar}}|$, the direct implementation requires $O(EN + dN^2)$ comparisons and $O(Nd + EN)$ stored input values.*

The quadratic term is only the naive pairwise dominance check. Faster Pareto routines can replace it for large grids without changing the decision rule. Missing target scores or missing cost coordinates narrow the claim rather than being silently imputed (Nie et al., 2023; Gulrajani & Lopez-Paz, 2021).

4 Information, Fitted Models, and Statistical Validity

Section 3 defines which configurations are admissible. A separate question is how that empirical decision should be interpreted. Three levels must be kept distinct: information available before learning, performance of the restricted fitted pipeline, and uncertainty in the estimated utility gaps.

4.1 ORACLE INFORMATION ORDER

Proposition 4 (Oracle monotonicity and information loss)**.** *Let $Z_1 = g(Z_2)$ almost surely for a measurable deterministic map. For any fixed decision loss and the class of all measurable decisions,*

$$\mathcal{R}^*(Z_2) \le \mathcal{R}^*(Z_1). \tag{9}$$

For a finite label space and log loss,

$$\mathcal{R}^*_{\log}(Z_1) - \mathcal{R}^*_{\log}(Z_2) = H(Y \mid Z_1) - H(Y \mid Z_2) = I(Y; Z_2 \mid Z_1) \ge 0. \tag{10}$$

Every decision based on Z_1 can be implemented from Z_2 by composition with g, so a deterministic coarsening cannot improve unrestricted oracle risk. In our setting this statement applies from the native observation to any deterministic coarsening, or between views for which a refinement relation is established. It does *not* imply that fitted macro-F1 must be monotone across independently resampled views or different representations.

4.2 WHY FITTED RANKINGS CAN REVERSE

For configuration $i \in \{1, 2\}$, let Z_i be its observation, $\mathcal{H}_i$ its hypothesis class, and $\widehat{h}_i \in \mathcal{H}_i$ the fitted predictor. Define

$$A_i = \inf_{h \in \mathcal{H}_i} \mathcal{R}(h; Z_i) - \mathcal{R}^*(Z_i) \ge 0, \qquad E_i = \mathcal{R}(\widehat{h}_i; Z_i) - \inf_{h \in \mathcal{H}_i} \mathcal{R}(h; Z_i) \ge 0.$$

The two terms separate restriction of the hypothesis class from imperfection of the fitted solution.

Proposition 5 (Fitted reversal decomposition)**.** *For each configuration,*

$$\mathcal{R}(\widehat{h}_i; Z_i) = \mathcal{R}^*(Z_i) + A_i + E_i. \tag{11}$$

If $Z_1 = g(Z_2)$ and $G_{\text{oracle}} = \mathcal{R}^*(Z_1) - \mathcal{R}^*(Z_2) \ge 0$, then the coarser fitted predictor outperforms the finer fitted predictor if and only if

$$(A_2 + E_2) - (A_1 + E_1) > G_{\text{oracle}}. \tag{12}$$

Equation (12) gives the precise reconciliation between information loss and a lower-rate fitted gain: the gain can occur only when the reduction in approximation-plus-fitting penalty exceeds the oracle information penalty. It does not identify which mechanism caused the observed wearable curves, because these terms are not estimated and macro-F1 is not the additive loss used in the decomposition.

A standard finite-sample sufficient bound gives one possible route. If $\sup_{h \in \mathcal{H}_i} |\widehat{\mathcal{R}}_i(h) - \mathcal{R}_i(h)| \le G_i$ and $\widehat{\mathcal{R}}_i(\widehat{h}_i) \le \inf_h \widehat{\mathcal{R}}_i(h) + \xi_i$, then

$$E_i \le 2G_i + \xi_i. \tag{13}$$

Reduced estimation or optimization difficulty can therefore change a fitted ranking without contradicting Proposition 4.

4.3 WHEN THE EMPIRICAL FRONTIER IS STABLE

Because the reference is itself selected by a maximum, uncertainty must control the whole candidate set rather than only the selected configuration.

Proposition 6 (Gap stability and set sandwich)**.** *Suppose*

$$\sup_{c \in \mathcal{C}} |\widehat{U}_e(c) - U_e(c)| \le \eta. \tag{14}$$

Then, for every $c \in \mathcal{C}$,

$$|\widehat{\Delta}_{e,\mathcal{C}}(c) - \Delta_{e,\mathcal{C}}(c)| \le 2\eta, \tag{15}$$

and

$$\widehat{\mathcal{S}}_{e,\mathcal{C}}(\epsilon - 2\eta) \subseteq \mathcal{S}_{e,\mathcal{C}}(\epsilon) \subseteq \widehat{\mathcal{S}}_{e,\mathcal{C}}(\epsilon + 2\eta), \tag{16}$$

where $\widehat{\mathcal{S}}(\alpha) = \varnothing$ for $\alpha < 0$.

The bound creates three regimes at the nominal tolerance: certified sufficient when $\widehat{\Delta}(c) + 2\eta \le \epsilon$, certified insufficient when $\widehat{\Delta}(c) - 2\eta > \epsilon$, and unresolved otherwise.

Corollary 1 (Frontier stability under a margin)**.** *Under Equation* (14)*, if*

$$|\widehat{\Delta}_{e,\mathcal{C}}(c) - \epsilon| > 2\eta \qquad \forall c \in \mathcal{C}, \tag{17}$$

then $\widehat{\mathcal{S}}_{e,\mathcal{C}}(\epsilon) = \mathcal{S}_{e,\mathcal{C}}(\epsilon)$*. If the measured cost vectors are exact, the empirical and population frontiers are identical as well.*

The same logic applies across targets. If $\sup_c |\widehat{U}_{s\to e}(c) - U_{s\to e}(c)| \le \eta_e$ simultaneously for every declared target, then

$$\max_{e \in \mathcal{E}_{\mathrm{tar}}(s)} \{\widehat{\Delta}_{s\to e}(c) + 2\eta_e\} \le \epsilon \tag{18}$$

certifies target-wise robust sufficiency; any absolute floor additionally requires $\widehat{U}_{s\to e}(c) - \eta_e \ge \tau_e$ for every target.

The theory therefore produces one coherent empirical program. Tolerance and reference changes should move the admissible set in predictable directions; stricter class, target, or perturbation requirements can invalidate an aggregate decision; and costs should affect preference only after utility preservation has been established. The next section evaluates those consequences with the available wearable results.

5 Experiments

5.1 Protocol

We evaluate the finite-grid decision rule on three public wearable benchmarks. The experiments are organized in the same order as the theory: first determine whether sufficiency changes the design decision, then test how that admissible set changes under stricter requirements, and finally examine cost preference within the surviving set.

Datasets and splits. UCI HAR records 30 participants at 50 Hz with an official participant-disjoint split over six activities (Anguita et al., 2013; Reyes-Ortiz et al., 2013). HAR70+ records 18 adults aged 70–95 at 50 Hz; we use the lower-back accelerometer and a four-class primary task (walking, sitting, standing, lying) with 2.56 s windows, 50% overlap, and a 0.8 label-purity threshold (Ustad et al., 2023; Logacjov & Ustad, 2023). Daphnet records ten participants with Parkinson's disease at 64 Hz; trunk acceleration is segmented into 4 s windows with 1 s stride for freezing-of-gait classification (Bächlin et al., 2010; Roggen et al., 2010). HAR70+ and Daphnet use three participant-group folds. The resolution intervention is controlled offline resampling rather than physical reacquisition.

Configurations and metric. UCI/HAR70+ use 1, 2, 5, 10, 25, and 50 Hz; Daphnet uses 1, 2, 4, 8, 16, 32, and 64 Hz. Polyphase anti-aliased resampling keeps physical window duration fixed. Four fixed representation families are compared: aggregate statistics, domain summaries, temporal/spectral statistics, and 64 frozen random convolutions. All use source-fitted standardization and a class-balanced Ridge probe. Macro-F1 is primary. We report $\epsilon = 0.01$ as the main benchmark tolerance and 0.02/0.03 as sensitivity values. The primary reference is the maximum over this fixed-feature grid; stronger released-feature results are treated only as contextual reference changes because their inputs and preprocessing differ. Appendix C gives the recoverable implementation details.

5.2 Sufficiency changes the design decision

The first question is whether the admissibility rule changes what would be selected from the same measured grid. It does on two of the three datasets. UCI's best tested point is the 10 Hz temporal/spectral configuration at 0.875 macro-F1, while the 5 Hz random-convolution configuration reaches 0.872, a gap of 0.003. HAR70+ remains at 10 Hz because its best 5 Hz score is 0.014 below the 0.872 reference. Daphnet's 64 Hz temporal/spectral reference is 0.619, while 32 Hz domain summaries reach 0.613, a gap of 0.006. Thus winner selection gives 10/10/64 Hz, whereas $\epsilon = 0.01$ sufficiency gives 5/10/32 Hz.

Table 1: Winner selection versus benchmark-relative sufficiency. "Best tested" identifies the configuration attaining the reference; the $\epsilon = .01$ column reports the best-scoring representation at the lowest sufficient tested rate.

Dataset	Best tested	U^{ref}	$\epsilon = .01$	Gap	.02	.03
UCI HAR	10 Hz / Temporal	0.875	5 Hz / Random conv.	.003	5 Hz	5 Hz
HAR70+	10 Hz / Temporal	0.872	10 Hz / Temporal	.000	5 Hz	5 Hz
Daphnet	64 Hz / Temporal	0.619	32 Hz / Domain	.006	16 Hz	8 Hz

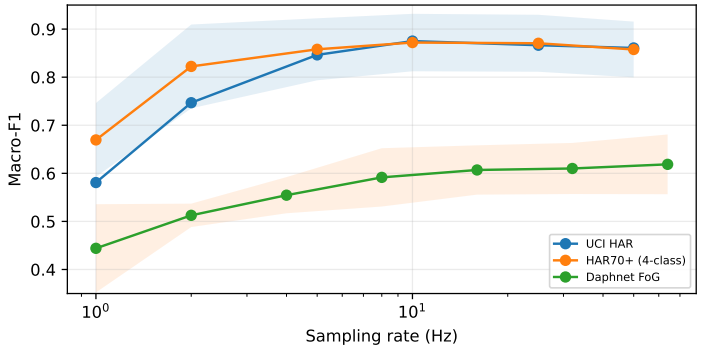


Figure 3: Resolution response for the temporal/spectral-statistics representation. HAR70+ and Daphnet show means and one standard deviation over three participant-group folds. Table 1 searches all four representation families.

The dependence predicted by Proposition 1 is visible without fitting any additional model. Relaxing the tolerance never increases the minimum sufficient tested rate: UCI stays at 5 Hz, HAR70+ moves from 10 to 5 Hz at $\epsilon = 0.02$, and Daphnet moves 32→16→8 Hz. Strengthening the reference moves the opposite way. On UCI, separately reported 345-feature SVM and 561-feature Ridge references reach 0.914 and 0.964, so the best compact-grid point at 0.875 would incur gaps of 0.039 and 0.089 under those stronger references. These heterogeneous feature views are not entries in the primary frontier; they demonstrate that sufficiency belongs to the declared comparison, not to a sampling rate in isolation.

Representation is part of that comparison. At native rate, UCI macro-F1 is 0.796/0.855/0.861/0.865 for aggregate/domain/temporal/random-convolution features, while Daphnet is 0.544/0.572/0.619/0.499. Yet UCI's lowest 0.01-sufficient point is supplied by random convolution rather than the temporal representation that defines the reference. Rate and representation therefore have to be selected jointly.

Figure 3 also contains a fitted reversal: UCI temporal/spectral statistics score 0.861 at 50 Hz and 0.875 at 10 Hz. Proposition 4 rules out the interpretation that discarding samples created new oracle label information; Proposition 5 shows how a restricted fitted pipeline can nevertheless reverse the oracle ordering.

5.3 Stricter requirements shrink or invalidate the aggregate decision

The within-cohort macro-F1 frontier is only one admissibility statement. Class-specific, target-cohort, and perturbation requirements can all be stricter. On UCI temporal features, walking F1 falls from 0.929 at 10 Hz to 0.660 at 2 Hz, and downstairs from 0.921 to 0.608, while lying remains nearly unchanged. Hence a small macro-F1 gap does not bound the worst class-specific loss; per-class tolerances would define a smaller admissible set. Detailed class curves and confusion matrices are moved to Appendix D so that the main text keeps the same decision logic rather than branching into a separate activity-analysis narrative. The same issue appears across environments. In magnitude-only four-class zero-shot transfer, UCI→HAR70+ temporal macro-F1 is 0.333 at 10 Hz and 0.358 at 25 Hz, while HAR70+→UCI is 0.395 and 0.496. Thus the 10 Hz configuration is at least 0.025 and 0.101 below a displayed alternative in the two directions, respectively, already violating the 0.01 target-wise tolerance. These are pairwise lower bounds on target regret rather than a complete

Table 2: Representation comparison at native rate (macro-F1). The $\pm$ terms are three-fold standard deviations, not uncertainty on pairwise score differences.

Dataset	Aggregate	Domain	Temporal	Random conv.
UCI HAR	0.796	0.855	0.861	0.865
HAR70+	0.834 ± 0.096	0.836 ± 0.094	0.858 ± 0.058	0.709 ± 0.048
Daphnet	0.544 ± 0.127	0.572 ± 0.092	0.619 ± 0.062	0.499 ± 0.040

Table 3: Representative $\epsilon = 0.01$ sufficient configurations and archived processing-cost measurements. Latency includes feature extraction and probe prediction on one CPU thread; probe bytes exclude the feature extractor and do not measure total deployment memory.

Dataset	Configuration	Macro-F1	Latency (ms)	Probe bytes
UCI HAR	Random conv., 5 Hz	0.872	3.40	7,376
UCI HAR	Temporal stats., 10 Hz	0.875	0.95	5,703
HAR70+	Temporal stats., 10 Hz	0.872	0.98	4,935
Daphnet	Domain summaries, 32 Hz	0.613	0.71	2,438

robust frontier, and the transfer combines hardware, placement, demographic, and protocol shift. The conclusion is therefore narrow but important: within-cohort sufficiency is not a property that can be transported across environments without re-evaluation. Data quality changes the same decision for a different reason. At UCI 10 Hz, temporal-statistics macro-F1 falls from 0.875 to 0.845 under Gaussian noise with standard deviation 0.02 and to 0.826 when 10% of time samples are replaced by the window mean; both losses exceed the primary tolerance. Reducing the number of training participants also lowers utility. These checks do not define a corruption-specific frontier because the reference is not recomputed for every perturbation, but they show that a clean-data admissible point need not remain admissible once the observation or training regime changes. Appendix D contains the full sensitivity curves and auxiliary transfer results.

5.4 Cost preference begins only after preservation

The final stage of the decision chain is cost comparison among configurations that already satisfy the utility requirement. Minimum sampling rate and minimum deployment cost need not coincide. On UCI, the 5 Hz random-convolution configuration uses fewer input samples but requires 3.40 ms median processing per window and a 7,376-byte probe; the 10 Hz temporal configuration requires 0.95 ms and a 5,703-byte probe. Neither dominates the other over the measured coordinates. The measured cost vector contains sampling rate, archived software processing latency, and serialized probe bytes. It does not contain sensor acquisition energy, filtering cost, radio use, power-state behavior, or device-specific memory. The current evidence therefore supports the ordering in Equation (5) only on the measured coordinates; direct hardware measurements would extend the same decision rule rather than create a different one.

6 Discussion and Conclusion

The main contribution is a change in what a temporal-resolution study treats as the object of decision. A best score identifies a winner; benchmark-relative sufficiency identifies the set of configurations whose utility remains acceptable, and only then asks which of them is preferable under deployment costs. The theory makes that object explicit by showing how the admissible set moves with tolerance, reference strength, target requirements, budgets, information loss, and estimation uncertainty. Across the tested wearable benchmarks, this distinction changes the selected rate from 10/10/64 Hz under winner selection to 5/10/32 Hz under a one-point macro-F1 tolerance. The resulting contribution is therefore not a universal sampling law. It is a formal way to turn a finite resolution sweep into an auditable design decision: specify what utility must be preserved, determine which tested configurations satisfy that requirement, and compare costs only within that set. Temporal resolution is sufficient only relative to that declared decision context.

AI USE STATEMENT

Generative AI (ChatGPT/Codex) assisted with literature retrieval, manuscript restructuring and language revision, refinement of definitions and mathematical arguments, interpretation of reported results, and preparation of the LaTeX and figure assets. This revision uses the previously reported empirical results; the AI-assisted editing workflow did not rerun the wearable experiments or generate new experimental observations. The mathematical additions and empirical interpretations require author verification before submission.

REPRODUCIBILITY STATEMENT

Sections 3–4 define the decision object, theoretical properties, and Algorithm 1; Section 5 gives the empirical instantiation. Appendix B contains the complete proofs, Appendix C documents the recoverable data-processing and representation details, and Appendices D–F provide diagnostics, algorithmic edge cases, and numerical checks. The supplied manuscript materials do not include the underlying experiment scripts, prediction files, or complete execution configuration, so the empirical measurements are documented but not independently reproducible from this source alone.

ETHICS STATEMENT

The study analyzes public wearable datasets, including older-adult and Parkinson's disease recordings. The reported relative sufficiency thresholds are benchmark results, not clinical validation or recommendations for patient monitoring. Cross-cohort failures and class-dependent errors motivate evaluation in the intended population and sensing conditions before deployment.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2013.

Marc Bächlin, Meir Plotnik, Daniel Roggen, Inbal Maidan, Jeffrey M. Hausdorff, Nir Giladi, and Gerhard Tröster. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):436–446, 2010.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lJkOCMP2aW.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Zichen Wang, Haoyang Hong, Chuanhao Li, Haoxuan Li, Zhiheng Zhang, and Huazheng Wang. Design-based bandits under network interference: Trade-off between regret and statistical inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Kadian Davis and Evans Owusu. Smartphone dataset for human activity recognition in ambient assisted living. UCI Machine Learning Repository, 2016.

Angus Dempster, François Petitjean, and Geoffrey I. Webb. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34:1454–1495, 2020. doi: 10.1007/s10618-020-00701-z.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1xwcyHFDr.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Simon Lee, Cyrus Tanade, Hao Zhou, Juhyeon Lee, Megha Thukral, Md. Sazzad Hissain Khan, Keum San Chun, Baiying Lu, Migyeong Gwak, Mehrab Bin Morshed, Viswam Nathan, Mahbubur Rahman, Li Zhu, Subramaniam Venkatraman, and Sharanya Desai. HiMAE: Hierarchical masked autoencoders discover resolution-specific structure in wearable time series. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6a3a5b1ad657c46ab99958b138bcd525-Abstract-Conference.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Aleksej Logacjov and Astrid Ustad. HAR70+. UCI Machine Learning Repository, 2023.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. Human activity recognition using smartphones. UCI Machine Learning Repository, 2013.

Daniel Roggen, Meir Plotnik, and Jeffrey M. Hausdorff. Daphnet freezing of gait. UCI Machine Learning Repository, 2010.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.08731.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=8qDwejCuCN.

Astrid Ustad, Aleksej Logacjov, Stine Overengen Trollebo, Pernille Thingstad, Beatrix Vereijken, Kerstin Bach, and Nina Skjæret Maroni. Validation of an activity type recognition model classifying daily physical behavior in older adults: The HAR70+ model. *Sensors*, 2023.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=PilZY3omXV2.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Maxwell A. Xu, Alexander Moreno, Hui Wei, Benjamin M. Marlin, and James M. Rehg. RETRIEVAL-BASED RECONSTRUCTION FOR TIME-SERIES CONTRASTIVE LEARNING. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=3zQo5oUvia.

Maxwell A. Xu, Jaya Narain, Gregory Darnell, Haraldur Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Fineman, Karthik J. Raghuram, James M. Rehg, and Shirley Ren. RelCon: Relative contrastive learning for a motion foundation model for wearable data. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k2uUeLCrQq.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with $10k$ parameters. In *International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=bWcnvZ3qMb.

A EVIDENCE SUMMARY AND EMPIRICAL SCOPE

This appendix collects the evidence behind the main empirical interpretations and states the scope in which each should be read. The purpose is not to introduce additional claims, but to make clear which observations in the paper support each conclusion and which extrapolations would require new experiments.

Table 4: Summary of the evidence supporting the main interpretations and the limits of that evidence.

Interpretation	Evidence used in this paper	Scope of the conclusion
Winner selection and sufficiency are different decisions	The same score grid evaluated with an explicit tolerance, reference set, and cost rule	The two decisions can select different rates; this does not compare the fixed features with all possible learned encoders
The tolerance controls the size of the admissible region	The $\epsilon = 0$ and large-ϵ limits, together with the $.01/.02/.03$ sensitivity table	The chosen values illustrate benchmark sensitivity; they are not universal or clinically validated margins
The sufficient rate is task dependent	UCI/HAR70+/Daphnet tolerance results	Different tasks admit different minimum tested rates; the experiment does not identify a continuous universal threshold
Rate and representation interact	Native-rate representation comparison and the UCI 5 Hz random-convolution / 10 Hz temporal-statistics results	Sufficiency depends on the pair (r, k); the evidence does not identify a causal reason why one feature family wins
Coarsening does not create oracle label information	Proposition 4 and the log-loss identity	The result constrains interpretation of score reversals; it does not identify the mechanism behind fitted macro-F1 changes
Aggregate macro-F1 can hide important failures	UCI class curves and the stronger released-feature references	The reported aggregate boundary depends on the metric and reference set; class-specific safety or clinical adequacy is not established
Within-cohort preservation need not transfer	UCI–HAR70+ and UCI–AAL comparisons	A source-cohort decision can fail in another cohort; the experiments do not constitute a complete robust frontier or isolate one source of shift
Clean-score sufficiency can be fragile	Noise/dropout and participant-subsampling results	Perturbation and data volume affect utility; full corruption-specific frontiers were not recomputed
Minimum rate need not minimize every measured cost	Selected latency and probe-byte measurements	The examples show a trade-off among measured costs; they do not establish battery savings, device latency, or global Pareto membership

Study setting. The acquisition-resolution question is posed for a fixed-duration wearable window: how much of the native temporal observation can be removed while preserving a declared downstream utility. This tension does not depend on a particular encoder, because additional temporal detail can increase acquisition and processing burden even when the utility gain is small. The empirical study uses real recorded datasets, but the rate intervention is controlled offline resampling and the tolerance is specified by the researcher. The results should therefore be interpreted as benchmark-relative preservation analyses rather than validation of deployed sensing policies.

Table 4 also makes clear that the main evidence is not simply the existence of a low-rate model with a high score. The informative comparison is the change in the design decision after a preservation threshold is imposed, together with the observed dependence on the reference, class, target cohort, perturbation, and measured costs.

B PROOFS AND INTERPRETATION

This section provides the proofs for Propositions 1–6, Theorem 1, and Corollary 1. The arguments are elementary but useful because they separate properties of the decision object from assumptions about any particular learning algorithm.

Proof of Proposition 1. If $c \in \mathcal{S}_{e,\mathcal{C}}(\epsilon_1)$ and $\epsilon_1 \le \epsilon_2$, then $\Delta_{e,\mathcal{C}}(c) \le \epsilon_1 \le \epsilon_2$, giving set inclusion. For $\mathcal{C} \subseteq \mathcal{C}'$, write $\delta = U^{\mathrm{ref}}_{e,\mathcal{C}'} - U^{\mathrm{ref}}_{e,\mathcal{C}} \ge 0$. For every shared c,

$$\Delta_{e,\mathcal{C}'}(c) = U^{\mathrm{ref}}_{e,\mathcal{C}'} - U_e(c) = \Delta_{e,\mathcal{C}}(c) + \delta,$$

which proves Equation (3); Equation (4) follows by thresholding both sides. Finally, on a finite grid the identity of the sufficient set changes only when ϵ crosses one of the finitely many gaps $\Delta_{e,\mathcal{C}}(c)$. Taking the minimum rate over nested sufficient sets yields a nonincreasing step function.

Proof of Proposition 2. If $\mathcal{C}(\mathbf{B})$ contains an ϵ-sufficient candidate c, then $V_e(\mathbf{B}) \ge U_e(c) \ge U^{\mathrm{ref}}_{e,\mathcal{C}} - \epsilon$. Conversely, if $U^{\mathrm{ref}}_{e,\mathcal{C}} - V_e(\mathbf{B}) \le \epsilon$, any budget-feasible maximizer of $V_e(\mathbf{B})$ is ϵ-sufficient. The equivalence follows.

Proof of Proposition 3. A candidate belongs to the target-wise intersection if and only if $\Delta_{s\to e,\mathcal{C}}(c) \le \epsilon$ for every target e, which is equivalent to the maximum target gap being at most ϵ. Increasing ϵ weakens all target inequalities; adding a target adds another conjunctive constraint and therefore cannot enlarge the set.

Proof of Theorem 1. The first loop computes the exact empirical reference for every declared target. The second loop retains a candidate exactly when its largest target-specific empirical gap is at most ϵ and all optional floors hold; by Proposition 3, this is precisely the empirical robust sufficient set. The nested dominance check removes exactly those sufficient candidates dominated in the declared cost coordinates, leaving Equation (5). The final minimum is the definition of the minimum sufficient tested rate. Computing all target gaps requires $O(EN)$ operations. The naive pairwise dominance check compares at most $N(N-1)$ pairs in d coordinates, giving $O(dN^2)$ time. The stated storage bound follows from the score matrix and cost vectors.

Proof of Proposition 4. Let $\mathcal{D}_i$ denote all measurable decision rules from Z_i to a common action space. If $Z_1 = g(Z_2)$, every $d_1 \in \mathcal{D}_1$ induces $d_2 = d_1 \circ g \in \mathcal{D}_2$ with identical actions almost surely, so taking infima gives $\mathcal{R}^*(Z_2) \le \mathcal{R}^*(Z_1)$. Under log loss, the infimum over all conditional predictive distributions equals $H(Y \mid Z)$. Because Z_1 is a function of Z_2, $H(Y \mid Z_1, Z_2) = H(Y \mid Z_2)$, and the chain rule gives Equation (10).

Proof of Proposition 5. By adding and subtracting the best risk in $\mathcal{H}_i$,

$$\mathcal{R}(\widehat{h}_i; Z_i) - \mathcal{R}^*(Z_i) = \left[\inf_{h\in\mathcal{H}_i} \mathcal{R}(h; Z_i) - \mathcal{R}^*(Z_i)\right] + \left[\mathcal{R}(\widehat{h}_i; Z_i) - \inf_{h\in\mathcal{H}_i} \mathcal{R}(h; Z_i)\right] = A_i + E_i.$$

Subtracting the decompositions for $i = 1, 2$ shows

$$\mathcal{R}(\widehat{h}_1; Z_1) - \mathcal{R}(\widehat{h}_2; Z_2) = G_{\mathrm{oracle}} + (A_1 + E_1) - (A_2 + E_2),$$

from which Equation (12) follows exactly. For Equation (13), upper-bound the population risk of $\widehat{h}_i$ by its empirical risk plus G_i, use approximate empirical optimality, and upper-bound the empirical risk of a population-optimal comparator in $\mathcal{H}_i$ by its population risk plus G_i.

Proof of Proposition 6. The elementary maximum inequality gives $|\max_c \widehat{U}_e(c) - \max_c U_e(c)| \le \eta$. Therefore

$$|\widehat{\Delta}(c) - \Delta(c)| \le |\widehat{U}^{\mathrm{ref}} - U^{\mathrm{ref}}| + |\widehat{U}(c) - U(c)| \le 2\eta.$$

If $\widehat{\Delta}(c) \le \epsilon - 2\eta$, then $\Delta(c) \le \epsilon$, proving the left inclusion in Equation (16). If $\Delta(c) \le \epsilon$, then $\widehat{\Delta}(c) \le \epsilon + 2\eta$, proving the right inclusion.

Proof of Corollary 1. If every empirical gap lies more than 2η from ϵ, each candidate is either certified sufficient or certified insufficient by Proposition 6; there is no ambiguous candidate. Hence the empirical and population sufficient sets coincide. With exact cost vectors, Pareto filtering the same finite set returns the same frontier.

These results do not provide a numerical confidence radius for the reported wearable experiments. They specify exactly what a valid radius would need to control and how such control would propagate from utility estimates to the final frontier.

C Experimental details and evidence availability

This section expands the protocol summarized in Section 5. It documents the implementation choices that can be recovered from the supplied materials, the source of the displayed measurements, and the execution details that remain unavailable. No missing value is imputed.

Representation definitions. For within-cohort triaxial acceleration $x_t \in \mathbb{R}^3$, features also use magnitude $m_t = \|x_t\|_2$. The aggregate-statistics representation computes mean, standard deviation, minimum, maximum, and RMS. The domain-summary representation adds mean absolute deviation, mean absolute first difference, signal-magnitude area, quartiles, median, and interquartile range. The temporal/spectral-statistics representation additionally uses skewness, excess kurtosis, centered zero-crossing rate, autocorrelation at lags 1, 2, and 4 when defined, spectral entropy, spectral centroid, and relative power in four Nyquist-relative bands. The random-convolution representation uses 64 frozen random kernels with lengths in $\{3, 5, 7, 9\}$ subject to available sequence length and returns the maximum and proportion of positive responses per kernel. The source-fitted standardization and class-balanced Ridge probe are common to all four families. Feature-count and serialized-extractor details are not recoverable from the supplied manuscript alone.

Physical support changes at coarse rates. At 1 Hz, a 2.56 s activity window contains only approximately three samples, depending on resampler rounding. A lag of four or a length-nine convolution is then unavailable. The original description limits such features to when they are defined, but does not specify all fallback rules. These rules, spectral estimators, kernel seeds/distributions, and resampling edge handling are needed to reproduce the exact pipeline. Similarly, autocorrelation at one sample corresponds to 20 ms at 50 Hz but 200 ms at 5 Hz. Thus a fixed family label does not imply fixed physical feature support.

Cross-cohort and auxiliary protocols. UCI–HAR70+ transfer uses only acceleration magnitude and four shared labels. Its comparisons are separated from six-class triaxial UCI evaluation. HAR70+ out-of-fold confusion counts give 96.9% walking recall and 93.2% lying recall at 10 Hz; 17.2% of sitting windows are classified as standing and 6.6% as lying. The UCI–AAL check uses released, name-aligned feature columns (Davis & Owusu, 2016). It does not resolve raw AAL label alignment. Gaussian perturbation levels are in the original signal units; dropout replaces selected time samples with the window mean. The noise injection order, seed, and complete evaluation code remain necessary execution details.

Source of the reported measurements. All displayed experimental measurements are retained from the supplied manuscript and PDF. Tables are transcribed from those values; original plots are recovered as vector crops, without estimating additional points from their appearance. The only new empirical arithmetic is explicit score differences and their regret lower bounds. The framework diagram is conceptual. The manuscript originally named the result files `uci_har_ridgecv_fast.csv`, `har70_common4_ridgecv_fast.csv`, `daphnet_ridgecv_fast.csv`, corresponding random-convolution files, and auxiliary transfer, noise, and sample-size files. Those files and experiment implementations were not among the three supplied attachments. This revision therefore does not assert that they are included or independently validated.

Details required for a complete empirical reproduction. The missing execution record includes participant-fold identities, Ridge hyperparameter grid and inner-validation grouping, all random seeds, raw feature-view selections, Daphnet exclusion/label-purity rules, exact resampler and feature defaults,

complete configuration-by-fold scores or predictions, and runtime hardware/software/measurement settings. These omissions prevent independent confirmation of a one-point statistical threshold and of global Pareto membership. No value for a missing setting has been fabricated in this package.

The available protocol is sufficient to interpret the displayed comparisons and their scope, but not to claim exact end-to-end reproducibility or statistical certification of the frontier.

D ADDITIONAL DIAGNOSTICS

The following diagnostics provide the class, transfer, corruption, and sample-size results referenced in Section 5.3. They supplement the primary sufficiency table by showing where its interpretation changes under different data conditions.

Table 5: Released-feature within-cohort references and UCI–AAL transfer. These feature views are outside the primary raw-accelerometer frontier.

Model / feature view	Train → test	Macro-F1	Bal. acc.
Ridge / 561 features	UCI → UCI	0.964	0.964
Ridge / 561 features	AAL → AAL	0.860	0.861
Ridge / 561 features	UCI → AAL	0.703	0.706
Ridge / 561 features	AAL → UCI	0.853	0.854
Linear SVM / 345 accel.	UCI → UCI	0.914	0.912
Linear SVM / 345 accel.	AAL → AAL	0.812	0.812
Linear SVM / 345 accel.	UCI → AAL	0.648	0.657
Linear SVM / 345 accel.	AAL → UCI	0.748	0.752

Table 6: UCI temporal-statistics perturbation results at 10 Hz. Dropout level is the fraction of time samples replaced by the window mean.

Perturbation	Level	Macro-F1
Clean	0.00	0.875
Gaussian noise	0.02	0.845
Gaussian noise	0.05	0.630
Gaussian noise	0.10	0.476
Sample dropout	0.05	0.848
Sample dropout	0.10	0.826
Sample dropout	0.20	0.756

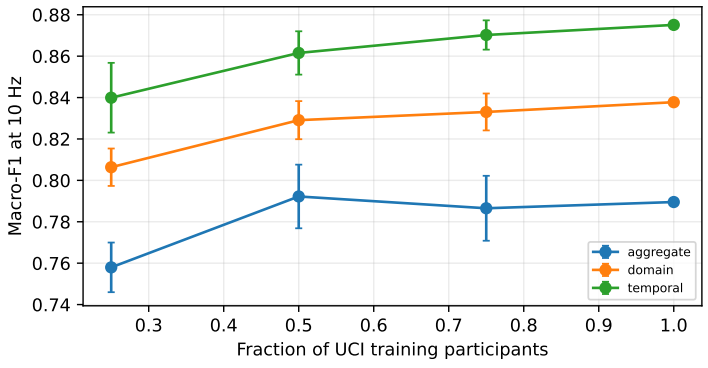


Figure 4: UCI participant-subsampling sensitivity at 10 Hz. Error bars use three seeds at reduced training fractions; the full-data point uses the official split. This fixed-rate analysis does not identify the mechanism of the resolution reversal.

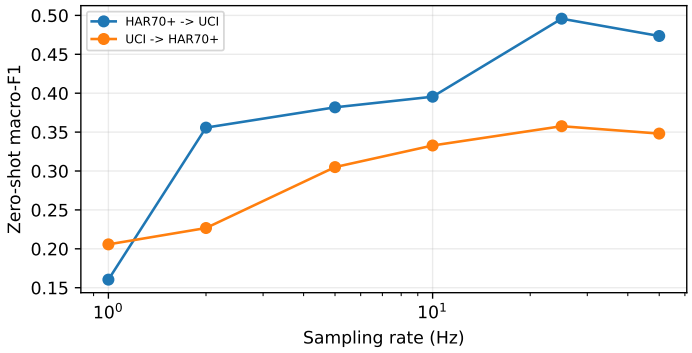


Figure 5: Full plotted magnitude-only, four-class UCI–HAR70+ transfer sweep preserved from the original study; the main text reports the exact displayed 10 Hz and 25 Hz values used for the target-wise comparison.

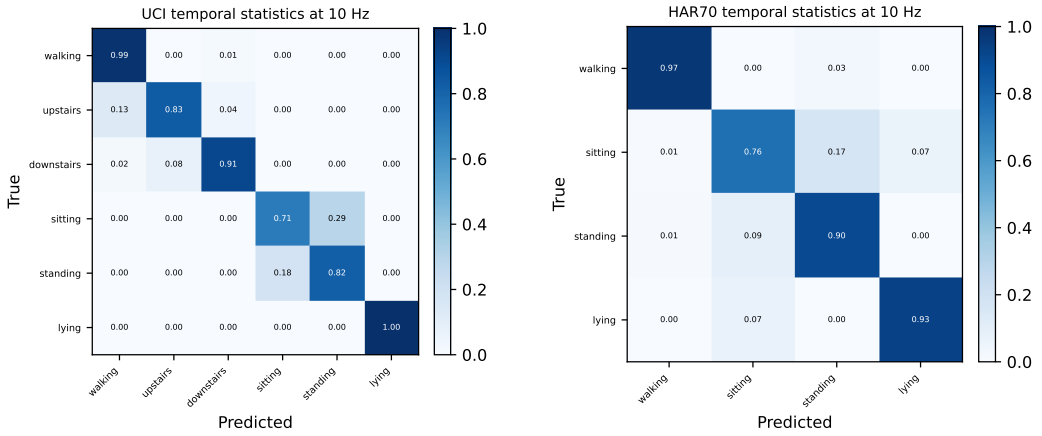


Figure 6: Row-normalized 10 Hz confusion matrices from the original study: UCI HAR (left) and four-class HAR70+ (right). Rounded cell labels may not sum to exactly one.

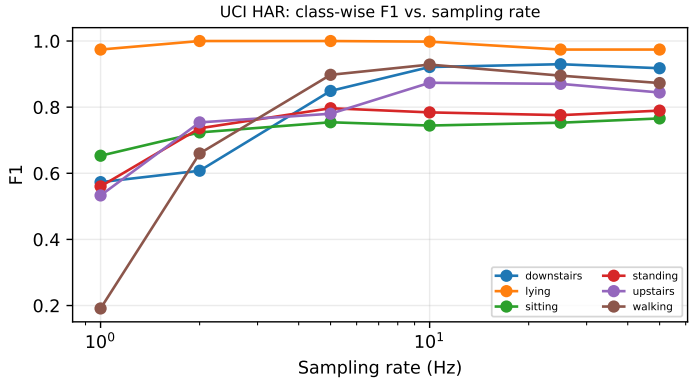


Figure 7: Class-specific sensitivity to input resolution on UCI HAR with temporal/spectral statistics. Dynamic activities lose substantially more performance under strong coarsening than lying, illustrating why aggregate macro-F1 sufficiency need not imply class-wise sufficiency.

Taken together, these diagnostics show that the clean, aggregate, within-cohort frontier is not automatically stable to class requirements, environment shift, corruption, or reduced training data. They do not identify a universal mechanism for those failures.

E FINITE-GRID ALGORITHM: EDGE CASES AND EXTENSIONS

Algorithm 1 is exact for a complete finite comparison table. This appendix records implementation details that matter when the table is incomplete or when uncertainty bounds are available.

Missing scores and costs. A candidate with a missing target score cannot be certified for the all-target intersection because its worst-target gap is undefined. A missing cost coordinate must not be silently replaced by zero; either obtain that measurement or report a frontier projected onto the coordinates measured for every compared candidate. Different projections can produce different dominance relations.

Equality and ties. A candidate with gap exactly ϵ is sufficient. Pareto dominance requires at least one strict cost improvement, so candidates with identical cost vectors remain tied unless an additional preference is declared in advance. Unrounded utilities should be used when available; rounding before thresholding can change membership for points near the boundary.

Uncertainty-aware filtering. Given simultaneous target-specific bounds η_e, a conservative robust set is

$$\widehat{\mathcal{S}}^{\mathrm{cert}}_{\mathrm{rob},s}(\epsilon) = \left\{ c : \max_{e \in \mathcal{E}_{\mathrm{tar}}(s)} [\widehat{\Delta}_{s \to e}(c) + 2\eta_e] \le \epsilon \right\}.$$

When absolute floors are part of the contract, also require $\widehat{U}_{s \to e}(c) - \eta_e \ge \tau_e$ for every target. These inequalities implement the sufficient side of Proposition 6; candidates in the ambiguity band should be reported as unresolved rather than forced into or out of the frontier.

Prospective selection. For a prospective study, $\mathcal{C}$, ϵ, target floors, and cost coordinates should be declared before final evaluation. Candidates can be fitted on source training data, the frontier selected on validation data, and the frozen decision evaluated on held-out participants or target environments. The current paper instead characterizes an archived finite grid retrospectively, so Algorithm 1 should be read as the exact computation implied by the evaluation object rather than as evidence that prospective selection was performed.

F METRIC DEFINITIONS AND NUMERICAL CHECKS

This section records the numerical details used by the main text: macro-F1 aggregation, the primary tolerance gaps, reference-set dependence, and the distinction between relative preservation and absolute utility. The calculations verify the reported arithmetic and interpretation; they do not add new experimental observations.

Macro-F1 and fold aggregation. For a declared label set of size K, write TP_j, FP_j, and FN_j for class-j confusion counts. Then

$$F_{1,j} = \frac{2\mathrm{TP}_j}{2\mathrm{TP}_j + \mathrm{FP}_j + \mathrm{FN}_j}, \qquad \mathrm{MacroF1} = \frac{1}{K} \sum_{j=1}^{K} F_{1,j}.$$

A zero denominator requires an explicit implementation convention. The supplied materials do not establish that convention for every fold; it should be recovered from the original evaluation implementation, rather than inferred. Averaging the three fold-level macro-F1 scores weights folds equally, whereas pooling their confusion counts generally gives a different result because F1 is nonlinear. The manuscript retains the original fold-level reporting convention.

Checking the primary gaps. The following arithmetic uses rounded values reported in the original manuscript. It makes their interpretation traceable without claiming additional statistical precision.

Comparison	Reference minus candidate	Gap	Interpretation at .01
UCI: 5 Hz random conv.	$0.875 - 0.872$	.003	Sufficient in reported grid
HAR70+: 5 Hz	$0.872 - 0.858$	.014	Insufficient
Daphnet: 32 Hz domain summaries	$0.619 - 0.613$	.006	Sufficient
Daphnet: 32 Hz temporal stats.	$0.619 - 0.610$	.009	Sufficient
UCI $\rightarrow$ HAR70+: 10 Hz temporal stats.	$0.358 - 0.333$	.025	Regret lower bound
HAR70+ $\rightarrow$ UCI: 10 Hz temporal stats.	$0.496 - 0.395$	.101	Regret lower bound

The last two rows compare 25 Hz with 10 Hz in the same transfer direction. If another tested candidate scores still higher, the true best-tested regret of 10 Hz is larger, so these differences remain valid lower bounds. The calculations do not require assuming that 25 Hz is globally best.

Reference-set dependence in concrete terms. The UCI compact-grid reference is 0.875, while the separately reported released-feature references reach 0.914 and 0.964. Substituting either as the utility target would create gaps of 0.039 and 0.089 even for the best compact configuration. These values explain why the stated compact-grid sufficiency cannot be described as preserving the strongest reported benchmark utility. Because the feature views and fitting procedures differ, they are contextual comparisons rather than controlled evidence about a single modeling component.

Why relative sufficiency alone is insufficient for deployment. As a purely illustrative example, if every candidate has target utility 0.30, then all candidates have zero target regret for any nonnegative tolerance, despite low absolute utility. Conversely, target-wise sufficient sets can be disjoint if one environment requires a configuration that another environment excludes. These logical possibilities motivate the intersection and utility-floor definitions. They are not additional wearable experiments.

Available supporting material. Appendices A–F provide the evidence summary, proofs, available protocol details, retained supplementary plots and tables, the explicit frontier procedure, and the numerical checks. The empirical plots and measurements are the supplied results. Because the underlying scripts and predictions are unavailable, no missing empirical value is reconstructed from figure appearance.

These checks support the interpretation used throughout the paper: sufficiency is relative to the declared metric, tolerance, reference set, environment, and measured costs. Stronger statistical or deployment conclusions require additional evidence rather than stronger wording.