USING RASHOMON SETS TO REDUCE IDENTITY LEAKAGE IN DATA COLLECTED FROM PEOPLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine Learning models increasingly recognize gestures and activities from behavioral data. The same measurements can also reveal who produced them. We use Rashomon sets and Model Class Reliance to construct models that retain the features most essential to a prediction task while simultaneously reducing reliance on identity-rich features. The models in a Rashomon set (i.e., all near-optimal models) often rely on different combinations of features. Treating identity as a second prediction task, we use Model Class Reliance to find features on which every near-optimal identity model depends, then trace effects of removing those features. We evaluate this approach on 10 public datasets from six sources, covering touch gestures, activity sensing and hospital records. As an example, in 14 of 18 ShearSense leave-one-subject-out folds, every near-optimal gesture tree uses at least one identity-reliant feature, so lower identifiability cannot be obtained simply by choosing another near-optimal gesture model. We therefore remove identity-reliant features one at a time, trace the full accuracy-identifiability curve, and cut it where further removal would cost more target task accuracy than identity accuracy. On the 6 datasets where the cut lowers identity accuracy by at least 5 percentage points, identity accuracy falls by 8.7 to 58.0 points and target task accuracy by 0.4 to 14.1 points. A pre-trained tabular deep model shows a similar target task cost. The tradeoff gives designers explicit control over target task accuracy and empirical identifiability, while identity retained in raw sensor data remains a system-design consideration.

1 Introduction

In some application spaces, machine learning models are trained on records of what people do. Phones and watches record movement, hospitals record encounters, schools record coursework. Many models built from these records are meant to serve the person recorded: wearable sensors track gait during stroke rehabilitation (Peters et al., 2021), keystroke timing indicates mental fatigue (Acien et al., 2022), a sensor-equipped steering wheel estimates a driver's stress (Cassani et al., 2021), and hospital records predict readmission (Strack et al., 2014). Here, we call the person behind a record a *contributor*, and whoever collects and models such records a *designer*. Because these task-directed measurements can often also reveal who produced them, this paper is concerned with the *adversary* who obtains and tries to use the features to recognize the contributor.

The Wireless Sensor Data Mining (WISDM) dataset, one of those examined here, was released for activity recognition and for biometric identification (Weiss et al., 2019). In domains from hospital discharges (Sweeney, 2002) to film ratings (Narayanan & Shmatikov, 2008), for decades, researchers have demonstrated that it is possible to identify individuals, using only a few attributes (Rocher et al., 2019) from records that have been stripped of names. Dropping the identifier column does not resolve the issue, because the remaining features carry an identity signal. A designer reading a feature list can neither tell which features carry an identity signal nor the cost of removing them. We study identity leakage wherever the data has a person-level identifier. Often, data collected from people include both a label for a *target task* (e.g. gesture recognition, patient readmission, activity recognition) the designer wants and an identity label for a potential *identity (ID) task* which the designer may have to limit, on behalf of the contributor, to meet a data-protection obligation. The procedure we propose offers the designer a way to *jointly manage* target task accuracy and ID

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.

leakage, and can be applied to any dataset which is subject to both tasks. We present results for nine activity-sensing datasets and one clinical set of hospital encounters.

Many datasets can be described with similar accuracy by models relying on different feature subsets (Breiman, 2001; Semenova et al., 2022). The models whose held-out loss on one task lies within ε of the optimal loss for that task comprise its *Rashomon set*. Here, we build separate Rashomon sets for target and ID tasks. While feature importance measures how significant a feature is to a *single* model, *model class reliance* (MCR) measures the importance of a feature in an entire set of models (Fisher et al., 2019). We use MCR of the Rashomon set to identify those features on which every near-optimal identity model depends, then ask whether any near-optimal target task model avoids them. When no such task model exists, we remove the identity-reliant features one at a time (in MCR order) and measure how much each removal costs in target task accuracy relative to the optimal (none left out) model.

This procedure runs before deployment. The designer uses person identifiers during an analysis step, chooses a *cut* (the number of identity-reliant features to remove, which fixes one point on the target task accuracy–identifiability curve), then decides which features downstream models can include. The best choice will depend on the context's need for conservatism: stronger reductions in identifiability can require larger losses in target task accuracy if no trained model in the Rashomon set is identity-independent. Differential privacy can then protect each deployed model as a second layer, because it bounds the risk of what a trained model reveals about the people in its training set.

To investigate this approach, we ask three questions of ten datasets:

RQ1 How much identity do the target task features carry, and which features does every near-optimal identity model rely on?

RQ2 How far can a different near-optimal target task model avoid the identity-reliant features, and how do target task accuracy and identifiability change as those features are removed?

RQ3 How well do these results hold using deep models, other feature-selection rules, a differentially private baseline, and different approximations of the Rashomon set?

We found that identity is predictable from the target task's own features on all 10 datasets (§5). Choosing a different near-optimal task model rarely avoids the features that every near-optimal identity model relies on, so reducing identifiability means removing features and paying for it in target task accuracy, which might be a feasible path in some application domains. In situations where the identity signal is concentrated in a small number of features, that exchange is favorable and it survives a deep adversary, five competing selection rules, and a differentially private baseline. Where many features carry overlapping identity information, few features are flagged, removing them changes both accuracies little, and identity accuracy stays well above chance.

2 Related work

This work connects four lines of research: identity prediction from behavioral data, re-identification from released records, model multiplicity, and methods that limit disclosure. In our setting, the adversary predicts person identity from the same features available to the target task model, so identity leakage and target task accuracy are measured on the same feature set.

Behavioral data and identity. Behavioral signals used for one prediction task can also support identity prediction. Epp et al. (2011) used keystroke timing to infer emotional state, and Kasprowski et al. (2022) used it to identify the typist. Wampfler et al. (2020) predicted affect from touch data, and Delgado-Santos et al. (2024) used touch data to authenticate users. Activity recognizers trained on wearable data can learn features that distinguish users (Iwasawa et al., 2017), and emotion recognizers trained on speech and text can encode demographic attributes (Jaiswal & Mower Provost, 2020). Machine learning has also re-identified adults and children from aggregated accelerometry data and demographics (Na et al., 2018), and a review of 72 wearable studies reported identification rates typically between 86 and 100% (Chikwetu et al., 2023). Wang et al. (2023) measure target task accuracy and re-identification accuracy on wearable sensor data, with an adversary trained on labeled samples of each person, and find that centering and scaling each person's features lowers re-identification accuracy and keeps or improves target task accuracy, and that a selected subset of 12

biomarker features lowers it further under this normalization. Our procedure instead orders feature removal using reliance across a Rashomon set and retrains the adversary on the features that remain.

Rashomon sets and model class reliance (MCR). Prior work extends feature analysis from one fitted model to sets of models with similar predictive performance. MCR reports the lowest and highest reliance on a feature across the near-optimal models of a class (Fisher et al., 2019). The Rashomon importance distribution (RID) represents importance as a distribution over the set (Donnelly et al., 2023). Semenova et al. (2022; 2023) and Rudin et al. (2024) studied set size and structure as properties of the learning problem. The TREEFARMS algorithm exactly enumerates Rashomon sets of sparse decision trees (Xin et al., 2022), the PRAXIS method approximates them with lower computational requirements (Heile et al., 2026a), and the ArborEnum method enumerates thresholds for continuous features without coarse binarization (Heile et al., 2026b). For sparse generalized additive models (GAMs), ellipsoids can approximate the Rashomon set (Zhong et al., 2023). Here, we sample many fitted trees and GAMs to approximate each Rashomon set (§3); App. B compares sets returned by TREEFARMS and PRAXIS, leaving ArborEnum to future work (§7).

Multiplicity and privacy. Model multiplicity lets one evaluate properties other than predictive accuracy across similarly performing models (Black et al., 2022). This was initially applied to fairness auditing Coston et al. (2021), then privacy-related properties of trained models. Differentially private training can produce equally private models that disagree on the same input (Kulynych et al., 2023). Hsu et al. (2025) evaluated enumerated Rashomon sets of sparse trees for criteria including resistance to membership inference. Hsu et al. (2026) showed that releasing multiple near-optimal models can expose more information about their training data than releasing one. These studies concern information about training records; we use the Rashomon set to identify input features associated with person identity and to test whether a near-optimal target task model can avoid them.

Other defenses. Differential privacy limits what a trained model reveals about its training records (Dwork & Roth, 2014), with methods for deep networks (Abadi et al., 2016) and random forests (Fletcher & Islam, 2017). Adversarial training instead learns representations that reduce information about a sensitive attribute (Edwards & Storkey, 2016; Ganin & Lempitsky, 2015), including user identity in activity data (Iwasawa et al., 2017; Malekzadeh et al., 2019) and demographic attributes in emotion recognition (Jaiswal & Mower Provost, 2020). Vertical data minimization restricts the information available to a deployed model by generalizing or removing feature values and evaluates reconstruction of client attributes (Staab et al., 2024). Here, we remove whole input features, retrain the identity adversary after removal, and use the target task accuracy–identifiability curve to show, for each possible cut, the target task accuracy lost and the identity removed.

3 Problem setup

Subjects, tasks and features. A dataset is a set of samples (x, y, u): features $x \in \mathbb{R}^p$, a target task label y, and a sample contributor u. In the evaluation we call the contributor the *subject*, as in leave-one-subject-out. The *target task* is to predict y; the *identity task* is to predict u from the same x. The procedure uses only x, y and u, so it applies to any dataset with a target task label and a person identifier: e.g., touch trials, sensor windows and hospital encounters.

Protocols. Table 1 lists the splits we use. We report mean $\pm$ SD across subjects for both tasks. We draw the held-out samples once and reuse the same split across all experimental arms (conditions being compared, each evaluated on the same subjects) and throughout the feature-removal sweep, so comparisons differ only in the features available to the model. Under this split, a model with no identity signal scores *below* $1/K$, because the held-out person contributes fewer training rows than the others. Identity accuracy below $1/K$ therefore means that the features carry no identity signal.

Threat model. Our threat model concerns access to the feature representation before model deployment. The adversary receives the same feature table that would be used to train or serve downstream models, together with labeled examples from individuals it wants to recognize. It trains its own identity classifier, and we measure how accurately it identifies held-out samples. The threat therefore comes from identity information retained in the shared or stored features, not from observing the deployed model itself. Attacks on a trained model are a separate threat. Differential privacy

Table 1: Evaluation splits. We use the first two to compute accuracies and the third to flag features. The target task Rashomon sets of Step 2 (§4) are fit inside each leave-one-subject-out (LOSO) fold.

Task	Split	Held out	Purpose
Target Task	leave-one-subject-out (LOSO)	Each subject in turn	Accuracy for people unseen in training
Identity Task	Per-subject adversarial	20% of each subject's samples	Accuracy with labeled examples of every person
Identity Rashomon sets	Stratified 5-fold over samples	Each fold decides membership; reliance uses 20% of training rows	Features every near-optimal identity model relies on (the flags)

addresses that setting by bounding what a trained model reveals about its training set; we include it as a baseline in §5.7.

Model reliance and MCR. Model reliance is the ratio of a model's loss with feature j permuted to its loss on the intact data,

$$\mathrm{MR}(j, m) = \frac{\mathbb{E}\left[\mathcal{L}\left(m, X^{(j)}\right)\right]}{\mathbb{E}\left[\mathcal{L}(m, X)\right]} \qquad \mathcal{L} = 1 - \text{accuracy} \tag{1}$$

estimated with 50 permutations of each feature, shared by every model in a fold's set. The Rashomon set $\mathcal{R}_\varepsilon$ holds every model whose held-out loss is within $\varepsilon = 0.05$ of the best model found, and for a given feature j, MCR_j^- and MCR_j^+ are the smallest and largest $\mathrm{MR}(j, \cdot)$ over it. $\mathrm{MCR}_j^- > 1$ means that every model in the set loses accuracy when j is permuted, and $\mathrm{MCR}_j^- \leq 1$ means that at least one model in the set does not rely on j. Sharing the permutations keeps MCR_j^- from picking up the luckiest permutations of any single model. We report the mean of MCR_j^- over folds. The split for the identity Rashomon sets (Table 1) differs from the adversarial split behind the accuracy numbers, so the two are not directly comparable.

How the Rashomon sets are built. We sample near-optimal models from two model classes. For trees, we fit a scikit-learn (Pedregosa et al., 2011) grid of 1,440 trees of depth 2 to 4, and for GAMs, we fit an interaction-free explainable boosting machines (EBMs) (Nori et al., 2019) at 24 boosting-round budgets. TREEFARMS and PRAXIS do not scale to most of our datasets and do not fit GAMs, so we use them only as a check (App. B). We add 4,000 trees and 200 EBMs with randomly drawn settings, which App. B lists. For each model class, we keep the fits whose performance lies within ε of the best of all its candidates. A grid alone yields few and similar near-optimal models, so a feature they all happen to use can look indispensable, and the varied candidates guard against that (App. B). On average, this leaves 80–215 GAMs and, depending on the dataset, 9 to 3,956 trees per fold. Throughout the paper, "every near-optimal model" means every model in these sampled sets.

4 Method

We conducted 12 experiments to address our research questions (§1): E1–E3 (RQ1), E4–E8 (RQ2), and E9–E12 (RQ3). Figure 4 (App. A) maps each step to its experiments. E1 and E7 evaluate all features and the features that remain after the cut with the splits of Table 1, E2 repeats identity evaluation with a blocked split (App. H), and E12 compares the Rashomon approximations (App. B).

4.1 The Rashomon-set procedure.

E3–E6 run Steps 1–4 on each dataset and record what each step finds.

Step 1: Flag the features every near-optimal identity model needs (E3). Fit Rashomon sets for the identity task in both model classes and flag feature j when $\mathrm{MCR}_j^- > 1.02$ in every fold for at least one model class, which means that permuting j raises the loss of even the least reliant near-optimal identity model of that class by more than 2% in every fold. Accepting either class is conservative, since it flags a feature that at least one of the two classes cannot do without. Requiring every fold, rather than the mean over folds, leaves out features whose reliance clears the threshold in some folds only (App. B).

Step 2: Filter best-performing models without the flagged features (E4). A tree can usually ignore any *one* feature, so the question for a deployment is whether one tree can ignore *all* the flagged features at once. In each fold we fit the target task Rashomon set on 80% of the training subjects, decide membership on the other 20%, and look for a member whose split features include no flagged feature. If one exists, the lowest-loss such member becomes the target task model and we refit nothing. If none exists, we refit a Rashomon set on the remaining features. Both routes run in every fold, so the two can be compared.

Step 3: Trace the curve (E5). Sort the flagged features by MCR^{-} and remove them one at a time, largest first, refitting after each removal and recording target task accuracy under leave-one-subject-out (LOSO) and identity accuracy under the adversarial split. Only flagged features enter this order, so the curve ends once all flagged features are removed. We use the same gradient-boosted model, with fixed settings, as the target task model and as the adversary at every step, so a change in the curve comes from the removed features.

Step 4: Choose the cut (E6). Record u_k for the share of above-chance identity accuracy that step k removes and v_k for the share of above-chance target task accuracy it costs; both run from 0 at the full feature set to 1 at chance, so datasets with different chance levels are comparable. The cut maximizes $u_k - v_k$, so removal continues while a step lowers identity accuracy more than target task accuracy, each as a share of its above-chance accuracy. The rule declines to cut when the whole sweep never lowers identity accuracy by at least 5 percentage points (pp), or when no step lowers identity accuracy more than target task accuracy, and we then remove every flagged feature. We choose each subject's cut on the curve averaged over the *other* subjects, so no subject's own score picks its own cut. Weighting identity and target task equally and requiring a reduction of 5 pp are our choices. App. E reports how the cut moves when the weight changes.

4.2 Evaluating the procedure (E8–E11) These experiments test the procedure's result with a person's own data (E8), a pretrained deep model (E9), five other selection rules (E10) and a differentially private forest (E11).

Personalization (E8). Deployments can recover accuracy with the user's own data. For each subject, we vary the fraction of that subject's samples that enter training from 0 to 1 and record target task accuracy on both the full and reduced feature sets.

A deep adversary and target task model (E9). We repeat both tasks, on the full and the reduced feature sets, with TabPFN v2 (Hollmann et al., 2025), a pretrained tabular model. We decompose tasks with more than ten classes one-vs-rest and cap the training context at 10,000 rows per fold, drawn stratified by label, which keeps every dataset inside the regime the model was pretrained for.

Comparisons (E10, E11). Using Rashomon sets to find and filter features is one possible approach. We compare it to five others at the same number of features (E10): mutual information with the target task, permutation importance of the best target task model, recursive feature elimination, principal components, and dropping features by one model's identity importance. We compare each approach with the flagged subset on both target task accuracy and identity accuracy. We also train a differentially private random forest (Fletcher & Islam, 2017) (E11) over a range of privacy budgets against a depth-matched non-private model.

Statistics. We compare methods subject by subject using a two-sided Wilcoxon signed-rank test (Wilcoxon, 1945). We treat all comparisons in one table, across all datasets, as a single family and apply Holm correction (Holm, 1979) to that family. We report the family size with each table. When we want to test whether a performance cost is small, we use two one-sided tests (TOST) with a ± 5 pp equivalence margin (Lakens et al., 2018). Datasets whose comparisons cannot reach significance after Holm correction are reported without tests and left out of the counts (App. C).

4.3 Datasets We analyzed 10 datasets from six public sources, summarized in Table 2 with details in App. A. ShearSense contains 8,295 touch-gesture trials from 18 participants, from which we compute 33 features. Eight activity datasets record wearable or phone sensors during daily and physical activities, with the activity as the target task, namely the four WISDM device and sensor pairs, PAMAP2, two OPPORTUNITY tracks (locomotion and gestures of both arms), and ExtraSensory. The clinical dataset, Diabetes 130-US Hospitals, records encounters of diabetes patients in 130 US hospitals, with readmission as the target task, which our baselines place near chance.

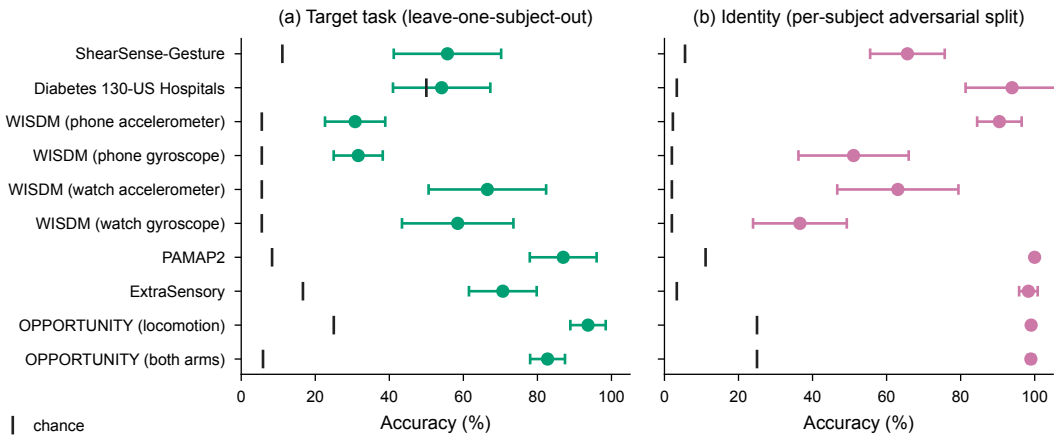


Figure 1: Target task features also support identity prediction. Accuracy of the gradient-boosted model (XGBoost; Chen & Guestrin, 2016) on all features, as mean ± SD across subjects.

5 Results

Features built for the target task identify people on every dataset (§5.1). On 9 datasets, the GAM Rashomon sets flag features that every near-optimal identity model relies on (§5.2). On 4 of them, most folds contain no near-optimal target task model that avoids those features (§5.3). On the 6 datasets where removing the flagged features lowers identity accuracy by at least 5 pp, identity accuracy falls by 8.7 to 58.0 pp and target task accuracy by 0.4 to 14.1 pp (§5.4, §5.5). On the 4 datasets with wide tables of windowed statistics, identity accuracy falls by at most 3.3 pp (§5.6). We repeat the comparisons with a pretrained deep model, five other feature-selection rules, a differentially private forest, and two other Rashomon-set constructions (§5.5, §5.7, §5.8).

5.1 Target task features support identity prediction on every dataset (E1, E2). With all features, identity accuracy ranges from $36.6 \pm 12.7\%$ on WISDM watch gyroscope, 18 times its 2.0% chance rate, to 98.3–100.0% on PAMAP2, both OPPORTUNITY tracks and ExtraSensory (Fig. 1). The tables contain no identifier column, so this identity signal comes from features built for the target task. With a blocked split that does not share related samples between training and test, identity accuracy is 97–99% on PAMAP2 and both OPPORTUNITY tracks (App. H).

5.2 The GAM Rashomon sets flag the identity-reliant features (E3). The flag rule marks 7 to 61% of the features on the 6 datasets with a few dozen engineered features, and at most 2% on the 4 wide tables of windowed statistics (Fig. 2). The GAM Rashomon sets flag every one of these features, and the tree sets flag a single feature, on OPPORTUNITY locomotion, which the GAM set also flags. A tree of depth at most 4 uses few features, so for almost every feature some near-optimal tree does not rely on it. A GAM adds one term per feature, so its identity accuracy builds up across many features (App. B).

5.3 Another near-optimal target task model avoids the flagged features on some datasets only (E4). On ShearSense and three of the four WISDM streams, most folds contain no near-optimal target task tree that avoids every flagged feature. On both OPPORTUNITY tracks, ExtraSensory and the clinical records, every fold contains one (Table 2). Where most folds contain none, lowering identifiability requires removing features.

5.4 The cut removes every flagged feature on all but one dataset (E5, E6). Removing the flagged features one at a time produces an accuracy–identifiability curve (Fig. 3). On every dataset except WISDM (watch accelerometer), the cut falls at the end of the curve. On WISDM (watch accelerometer), it falls one step before the end. Choosing each subject's cut from the other subjects' curves gives the same cut for 237 of 242 subjects on the datasets with a knee (Table 12).

5.5 Where the cut removes features, identity accuracy falls by more than target task accuracy (E7, E8, E9). On the 6 datasets where the cut lowers identity accuracy by at least 5 pp, identity

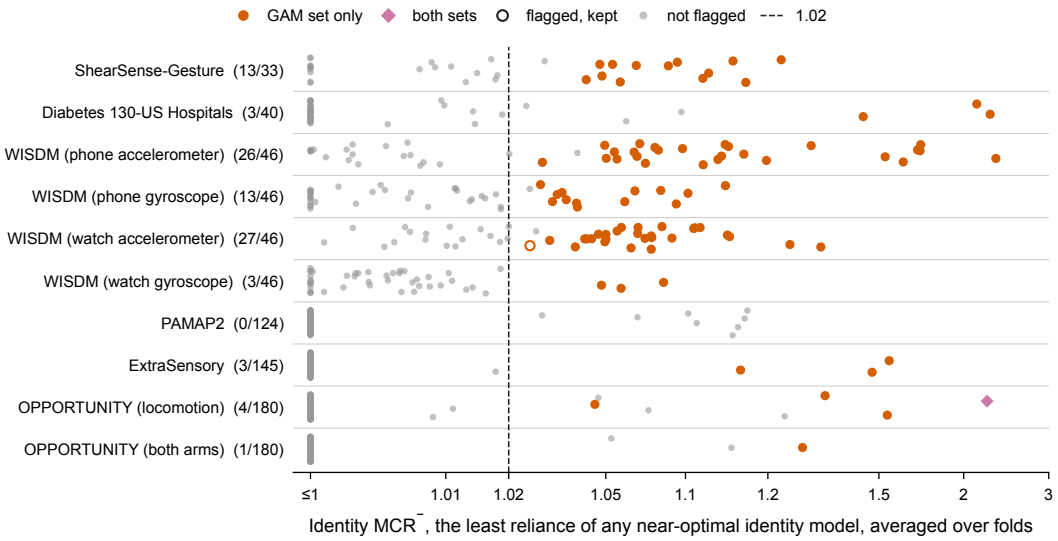


Figure 2: Features used across near-optimal identity models. On 8 of the 10 datasets, the GAM set alone flags every flagged feature. Each point is a feature at its fold-mean MCR^{-}, the larger of the tree and GAM values. A flagged feature exceeds 1.02 in every fold, so gray points past the line fall short of it in some fold. Filled points are removed at the cut, and parentheses give the number removed out of the total. The x-axis is logarithmic in $\mathrm{MCR}^{-} - 0.995$.

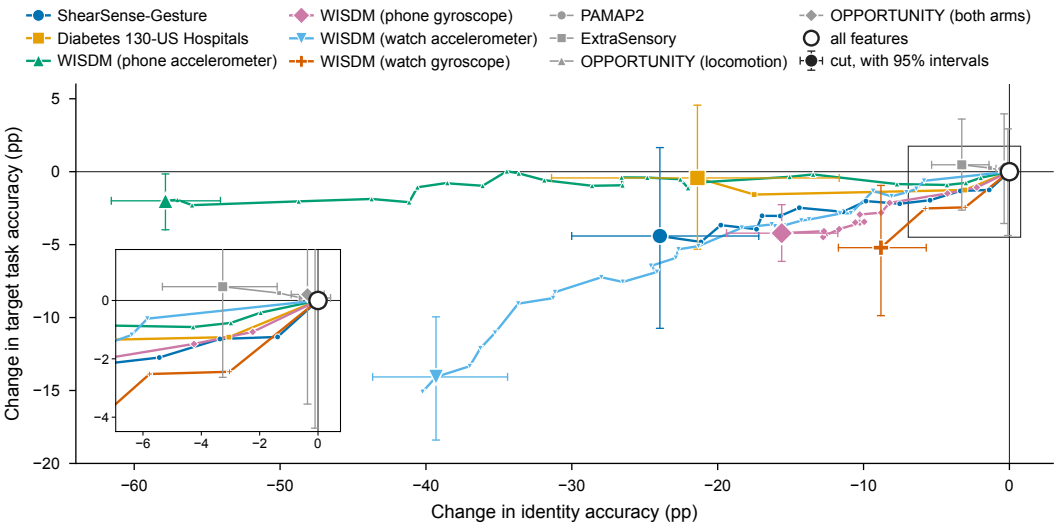


Figure 3: Cuts on the accuracy–identifiability curves. At every colored cut, identity accuracy falls by more than target task accuracy. The gray curves of the high-dimensional sets change little, and even in the inset the ring hides the OPPORTUNITY locomotion curve, which stays within 0.2 pp of the origin. Each curve removes one dataset's flagged features one at a time. Both axes show the change from all features in the gradient-boosted model's mean accuracy across subjects, on one scale. Crosses span 95% bootstrap intervals over subjects. The inset, enlarges the boxed region at the origin twofold and cuts intervals off at its edge.

accuracy falls by 8.7 to 58.0 pp and target task accuracy by 0.4 to 14.1 pp. The largest identity drop, 58.0 pp on WISDM phone accelerometer, comes with a target task change of -2.0 pp, which an equivalence test places within ± 5 pp (Table 6). After the cut, identity accuracy on these datasets is 7 to 22 times chance. With TabPFN, a pretrained tabular model, identity accuracy also falls by at least 5 pp on all 6 of them. Trained with a person's own data, the retained features give higher target task accuracy than the all-feature population model on 6 of the 7 datasets where identity accuracy decreases significantly (Table 9).

5.6 Where identity information is spread across substitutable features, removal lowers identity accuracy by at most 3.3 pp (E5). On PAMAP2, both OPPORTUNITY tracks and ExtraSensory, identity accuracy falls by at most 3.3 pp at the cut, and PAMAP2 has no flagged features. These datasets are wide tables of windowed statistics. When another feature can substitute for feature j, a near-optimal identity model can rely on that substitute, which keeps MCR_j^- near 1 and leaves j unflagged.

5.7 Other selection rules at the same number of features have higher identity accuracy (E10, E11). At the same number of features, none of five other selection rules gives both significantly higher target task accuracy and significantly lower identity accuracy than the flagged subset. In 26 of 33 comparisons, the other rule has significantly higher identity accuracy (App. C). Dropping features by one model's identity importance selects similar features (median Jaccard 0.90) and differs significantly from our procedure on 2 of 5 datasets. The Rashomon-set procedure also tests whether a near-optimal target task model avoids the flagged features (§5.3). A differentially private forest at $\varepsilon = 1$ has identity accuracy below chance on all 10 datasets and lowers target task accuracy by 10.3 to 61.2 pp on 7 of them (Table 11). Its guarantee bounds what the model reveals about its training set.

5.8 The grid-only sets flag every feature our sets flag (E12). We compare our Rashomon sets with sets sampled from the grid alone, and our tree sets with those that TREEFARMS and PRAXIS enumerate (App. B). On 10 of the 10 datasets, every feature our sets flag is also flagged by the grid-only sets, which flag more features on most datasets (Table 4). In the enumerated tree sets, no feature has $\mathrm{MCR}^- > 1.02$ on ShearSense, WISDM watch accelerometer or the clinical records, as in our sampled tree sets. On 8 of the 10 datasets, the GAM Rashomon sets alone produce every flag (App. B).

6 Discussion

Where feature removal changes identifiability. On ShearSense, the four WISDM streams, and the clinical records, removing identity-reliant features reduces identity accuracy at a target task cost of at most 14.1 pp (§5.5). On PAMAP2, both OPPORTUNITY tracks, and ExtraSensory, which hold wide tables of windowed statistics with substitutable features, removing the flagged features changes identity accuracy little (§5.6). For these wide tables of windowed statistics, reducing identity might require changing the representation itself (Malekzadeh et al., 2019) or restricting which sensor signals are available downstream.

Role of the Rashomon set. The conclusions of the Rashomon analysis apply to the sampled model sets. The varied candidates remove flags that a grid-only sample produces, and every flag they keep was already a grid flag (§5.8). For feature selection alone, one identity model produces a similar subset, with median Jaccard 0.90, and is statistically indistinguishable from our procedure on both tasks in 3 of 5 testable datasets (§5.7). The Rashomon set additionally tests whether any near-optimal target task model avoids all flagged features.

Residual identity and the accuracy cost. Identity accuracy remains above chance after feature removal. On datasets where identity falls by at least 5 pp, it remains between 7 and 22 times chance (§5.5), so the retained features do not make the data anonymous. Trained with the person's own data, the retained features give higher target task accuracy than the all-feature population model on 6/7 datasets where identity accuracy decreases significantly (App. I). App. I discusses the accuracy cost in more detail, including low-cost cuts, alternative protections, and relevant EU regulatory context.

Limitations. With 9 subjects in PAMAP2 and 4 in OPPORTUNITY, no OPPORTUNITY comparison, and no PAMAP2 comparison at the cut or against the selection rules, can reach significance after Holm correction, so we report those results without significance tests. We study engineered features, and TabPFN tests a stronger learner on the same representation but does not test sequence models trained on raw inputs. The Rashomon sets cover trees and EBMs and are sampled, so the flags describe every near-optimal model we found, not every near-optimal model, and a differently built set can flag different features (App. B). By analogy with the privacy onion effect (Carlini et al., 2022), recomputing the flags after each removal could identify substitutes that become important

later, since we compute them once from the full feature set. The LOSO folds share training subjects, so paired tests over folds might underestimate uncertainty (Dietterich, 1998; Nadeau & Bengio, 2003). Feature selection also uses samples later held out by the identity adversary, so a nested evaluation would give a stricter estimate of the reduction. Finally, the PAMAP2 and OPPORTUNITY preprocessing choices are specific to our pipeline (App. A), and we do not evaluate cross-session identification.

7 Conclusion and future work

We asked which features in data collected from people support identity prediction and how removing those features changes target task accuracy. On all 10 datasets, the target task features support identity prediction above chance. For datasets with a few dozen engineered features (ShearSense, WISDM and the clinical records), identity reliance is concentrated in a subset of features, and removing those features can reduce identity accuracy more than target task accuracy. For wider tables of windowed statistics, removing individually flagged features produces smaller changes, because identity information is distributed across substitutable features. Identity accuracy remains above chance after feature removal on every dataset.

Designers with access to person-level identifiers can run the procedure prior to releasing features or training deployed models, to learn feature identity content, degree to which a near-optimal target task model avoids identity-reliant features, and the cost in target task accuracy of removing them. This can inform consent materials, documentation, and data-protection risk assessments (App. I).

Several extensions follow from the limitations of this study. Selecting features inside a nested split would evaluate identity reduction on samples that were not used to choose the features. Measuring reliance over groups of correlated features could address identity information distributed across substitutable features, which would extend the procedure to the wide tables of windowed statistics where it now has limited effect. The Rashomon approximations could also be extended, using exact enumeration over continuous features for trees (Heile et al., 2026b) and an ellipsoid approximation for GAMs (Zhong et al., 2023), to test whether the flags hold beyond the sampled sets. Future comparisons could include per-person centering and scaling (Wang et al., 2023), a baseline that transforms the features instead of removing them. Cross-session identity evaluation and sequence models trained on raw signals would test settings not covered by the present protocol, including an adversary that recognizes a person on a later day or from the raw signal. We release code, features, and analysis scripts to support replication and extension.

AI use statement

We used Claude Code (Anthropic) to write the analysis pipeline, the statistics and digest scripts, the figure scripts, and the script that generates every number in this paper from the archived results. It also drafted outlines and passages of the text based on the authors' own earlier drafts, and the authors rewrote and edited them extensively. We checked the code with unit tests and by reproducing archived values and a separate verification pass recomputed the numbers in the paper from the archived per-subject tables. The authors checked the citations against their sources and take responsibility for all content.

Ethics statement

Every dataset here is public and was collected by others, under the consent and ethics approval of the original study. We collected no new data and did not contact participants. We checked the terms each dataset is released under before running an identity task on it, since that task is not the purpose the data was gathered for. PAMAP2, OPPORTUNITY, WISDM and Diabetes 130-US Hospitals are released under CC BY 4.0, which permits any use with attribution, and WISDM's own documentation states that the data supports behavioral biometric models, because every sensor reading carries its subject. The terms under which ShearSense and ExtraSensory are released also permit this analysis, which we confirmed for each release.

We identify people only within a dataset, from the subjects the dataset already labels, and we link nothing to any outside source or to any real name. The reported accuracies measure how much identity a feature set carries. In other words, they are an audit of a risk that a deployment would

otherwise carry unknowingly. The same measurement points to the features an attacker would find most useful, which is the cost of publishing this kind of audit, and we judge the cost worth paying because a designer cannot manage a risk that nobody has measured. We report the reductions we measured. A limitation of our work is that it is not a blanket safety mechanism, which means that a different adversary, or the raw signal behind the features, might recover more identity than we measured.

Reproducibility statement

We developed a single pipeline that produces every result in this paper. It runs in phases, from the feature table through the baselines, the Rashomon sets and MCR, the removal sweep, the deep model, the selection rules, the private baseline and the personalization curve, and it is driven per dataset from a single entry point. Everything a dataset means lives in its adapter, which names the column that identifies the person, the target task label, the feature columns and the artifact screens that apply. No experimental phase was developed specifically for a single dataset, so the same code runs on touch trials, sensor windows and hospital encounters, and adding a dataset means writing a new adapter and leaving every phase as it is.

We lock dependencies to exact versions, and the environment file, both container images and the cluster recipe carry the same pins. The deep tabular model pins separately, because it requires older versions of two libraries. Each run was done with fixed seeds. The phases that fit models record the dataset, the artifact screen, the seed, the resolved library versions and that phase's configuration. The results this paper reports come from one archived run, kept read only, whose provenance file records the code and the approximation it used. The code changed during that run only to add the two datasets we included last and to let one phase resume where it stopped, and a check against the changes shows that they leave the other eight datasets' results as they were.

The statistics are computed once, in one place. Each phase stores a score for every subject. A p-value here never depends on which phase produced the experimental arm. We also fix the family for each table and apply Holm across every comparison in it.

We use a script to get every number in this text from the digest and the archived, read-only tables and write a LaTeX macro for each number, with a comment naming the file, row and column it came from; an index of the keys ships with it, and its own record stores the checksum of the digest tables it read. If a specific key does not resolve, it prints in red in the compiled paper, so a renamed comparison is visible. Table 2 and the counts in §4.3 are generated the same way.

We implemented a test suite covering the evaluation protocols, the statistics, the selection rules, the dataset adapters and the number generator, and the tests print the expected and computed values side by side so a reader can check them. It passes, with one known failure in a feature-level edge case that the analyses do not rely on, since they read the cached feature tables. An independent pass re-derived every number and method statement in an earlier draft from the archived tables.

The supplementary material holds the code, the pins, the container recipes, the archived per-subject tables and the scripts that rebuild the digest, the figures and the numbers. Dataset construction is described in App. A, how the Rashomon sets were built and checked in App. B, and the statistical procedure in §4.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*, pp. 308–318. ACM, 2016. doi: 10.1145/2976749.2978318.

Alejandro Acien, Aythami Morales, Ruben Vera-Rodriguez, Julian Fierrez, Ijah Mondesire-Crump, and Teresa Arroyo-Gallego. Detection of mental fatigue in the general population: Feasibility study of keystroke dynamics as a real-world biomarker. *JMIR Biomedical Engineering*, 7(2): e41003, 2022. ISSN 2561-3278. doi: 10.2196/41003.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. Window size impact in human activity recognition. *Sensors*, 14(4):6474–6499, 2014. ISSN 1424-8220. doi: 10.3390/s140406474. URL https://www.mdpi.com/1424-8220/14/4/6474.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pp. 850–863. ACM, 2022. doi: 10.1145/3531146.3533149.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001. doi: 10.1214/ss/1009213726.

Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.*, 46(3), January 2014. ISSN 0360-0300. doi: 10.1145/2499621. URL https://doi.org/10.1145/2499621.

Xi Laura Cang, Paul Bucci, Andrew Strang, Jeff Allen, Karon E. Maclean, and H. Y. Sean Liu. Different strokes and different folks: Economical dynamic surface sensing and affect-related touch recognition. In *Proceedings of the 2015 ACM International Conference on Multimodal Interaction (ICMI '15)*. ACM, 2015. doi: 10.1145/2818346.2820756.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramèr. The privacy onion effect: Memorization is relative. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13263–13276, 2022. doi: 10.52202/068431-0964.

Raymundo Cassani, Atsushi Horai, Lucian A. Gheorghe, and Tiago H. Falk. Predicting driver stress levels with a sensor-equipped steering wheel and a quality-aware heart rate measurement algorithm. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 6822–6825, Mexico, 2021. IEEE. ISBN 978-1-7281-1179-7. doi: 10.1109/EMBC46164.2021.9630951. URL https://ieeexplore.ieee.org/document/9630951/.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Lucy Chikwetu, Yu Miao, Melat K. Woldetensae, Diarra Bell, Daniel M. Goldenholz, and Jessilyn Dunn. Does deidentification of data from wearable devices give us a false sense of security? A systematic review. *The Lancet Digital Health*, 5(4):e239–e247, 2023. doi: 10.1016/S2589-7500(22)00234-5.

John Clore, Krzysztof Cios, Jon DeShazo, and Beata Strack. Diabetes 130-US hospitals for years 1999-2008. UCI Machine Learning Repository, 2014.

Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2144–2155. PMLR, 2021.

I. Croy, A. Luong, C. Triscoli, E. Hofmann, H. Olausson, and U. Sailer. Interpersonal stroking touch is targeted to c-tactile afferent activation. *Behavioural Brain Research*, 297:37–40, 2016. doi: 10.1016/j.bbr.2015.09.038.

Paula Delgado-Santos, Ruben Tolosana, Richard Guest, Parker Lamb, Andrei Khmelnitsky, Colm Coughlan, and Julian Fierrez. Swipeformer: Transformers for mobile touchscreen biometrics. *Expert Systems with Applications*, 237:121537, 2024. ISSN 09574174. doi: 10.1016/j.eswa.2023.121537.

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998. doi: 10.1162/089976698300017197.

Jon Donnelly, Srikar Katta, Cynthia Rudin, and Edward P. Browne. The Rashomon importance distribution: Getting RID of unstable, single model-based variable importance. In *Advances in Neural Information Processing Systems*, pp. 6267–6279. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-0274.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/0400000042.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations (ICLR)*, 2016. doi: 10.48550/arXiv.1511.05897.

Clayton Epp, Michael Lippold, and Regan L. Mandryk. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, pp. 715–724. ACM, 2011. doi: 10.1145/1978942.1979046.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88, 2016. URL `http://data.europa.eu/eli/reg/2016/679/oj`.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, 2024. URL `https://eur-lex.europa.eu/eli/reg/2024/1689/oj`.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Sam Fletcher and Md Zahidul Islam. Differentially private random decision forests using smooth sensitivity. *Expert Systems with Applications*, 78:16–31, 2017. doi: 10.1016/j.eswa.2017.01.034.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189. PMLR, 2015.

Steven C. Hauser, S. McIntyre, A. Israr, H. Olausson, and G. J. Gerling. Uncovering human-to-human physical interactions that underlie emotional and affective touch communication. In *World Haptics Conference*, pp. 407–412. IEEE, 2019. doi: 10.1109/WHC.2019.8816169.

Zakk Heile, Hayden McTavish, Varun Babbar, Margo Seltzer, and Cynthia Rudin. From Rashomon theory to PRAXIS: Efficient decision tree Rashomon sets. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026a. doi: 10.48550/arXiv.2606.00202. arXiv:2606.00202.

Zakk Heile, Hayden McTavish, Margo Seltzer, and Cynthia Rudin. ArborEnum: Decision tree Rashomon sets over continuous features. *arXiv preprint arXiv:2608.04310*, 2026b.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Ethan Hsu, Tony Cao, Lesia Semenova, and Chudi Zhong. The Rashomon set has it all: Analyzing trustworthiness of trees under multiplicity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pp. 152738–152785, 2025. doi: 10.52202/085713-4596.

Ethan Hsu, Harry Chen, Chudi Zhong, and Lesia Semenova. The double-edged nature of the Rashomon set for trustworthy machine learning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. doi: 10.48550/arXiv.2511.21799. arXiv:2511.21799.

Yusuke Iwasawa, Kotaro Nakayama, Ikuko Yairi, and Yutaka Matsuo. Privacy issues regarding the application of DNNs to activity-recognition using wearables and its countermeasures by use of adversarial training. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17)*, pp. 1930–1936. International Joint Conferences on Artificial Intelligence Organization, 2017. doi: 10.24963/ijcai.2017/268.

Mimansa Jaiswal and Emily Mower Provost. Privacy enhanced multimodal neural representations for emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 7985–7993, 2020. doi: 10.1609/aaai.v34i05.6307.

Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 1895–1912, Santa Clara, CA, 2019. USENIX Association. URL https://www.usenix.org/conference/usenixsecurity19/presentation/jayaraman.

Pawel Kasprowski, Zaneta Borowska, and Katarzyna Harezlak. Biometric identification based on keystroke dynamics. *Sensors*, 22(9):3158, 2022. ISSN 1424-8220. doi: 10.3390/s22093158.

Bogdan Kulynych, Hsiang Hsu, Carmela Troncoso, and Flavio du Pin Calmon. Arbitrary decisions are a hidden cost of differentially private training. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, pp. 1609–1623. ACM, 2023. doi: 10.1145/3593013.3594103.

Daniël Lakens, Anne M. Scheel, and Peder M. Isager. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269, 2018. doi: 10.1177/2515245918770963.

Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation (IoTDI '19)*, pp. 49–58. ACM, 2019. doi: 10.1145/3302505.3310068.

Devyani McLaren, Jian Gao, Xiulun Yin, Rúbia Reis Guerra, Preeti Vyas, Chrys Morton, Xi Laura Cang, Yizhong Chen, Yiyuan Sun, Ying Li, John David Wyndham Madden, and Karon E. MacLean. What is affective touch made of? A soft capacitive sensor array reveals the interplay between shear, normal stress and individuality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*, pp. 52:1–52:31. ACM, 2024. doi: 10.1145/3654777.3676346.

India Morrison, Line S. Löken, and Håkan Olausson. The skin as a social organ. *Experimental Brain Research*, 204(3):305–314, 2010. doi: 10.1007/s00221-009-2007-y.

Liangyuan Na, Cong Yang, Chi-Cheng Lo, Fangyuan Zhao, Yoshimi Fukuoka, and Anil Aswani. Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. *JAMA Network Open*, 1(8):e186040, 2018. doi: 10.1001/jamanetworkopen.2018.6040.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52 (3):239–281, 2003. doi: 10.1023/a:1024068626366.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy*, pp. 111–125. IEEE, 2008. doi: 10.1109/SP.2008.33.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. doi: 10.5555/1953048.2078195.

Denise M. Peters, Emma O'Brien, Kira E. Kamrud, Shawn M. Roberts, Talia A. Rooney, Kristen P. Thibodeau, Swapna Balakrishnan, Nancy M. Gell, and S. Mohapatra. Utilization of wearable technology to assess gait and mobility post-stroke: a systematic review. *Journal of NeuroEngineering and Rehabilitation*, 18(1):67, 2021. doi: 10.1186/s12984-021-00863-x.

Attila Reiss. PAMAP2 physical activity monitoring. UCI Machine Learning Repository, 2012.

Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1): 3069, 2019. doi: 10.1038/s41467-019-10933-3.

Daniel Roggen, Alberto Calatroni, Long-Van Nguyen-Dinh, Ricardo Chavarriaga, and Hesam Sagha. OPPORTUNITY activity recognition. UCI Machine Learning Repository, 2010.

Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Position: Amazing things come from having many good models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42783–42795. PMLR, 2024.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pp. 1827–1858. ACM, 2022. doi: 10.1145/3531146.3533232.

Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. A path to simpler models starts with noise. In *Advances in Neural Information Processing Systems 36*, volume 36, pp. 3362–3401. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-0149.

Robin Staab, Nikola Jovanović, Mislav Balunović, and Martin Vechev. From principle to practice: Vertical data minimization for machine learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 4733–4752, 2024. doi: 10.1109/SP54263.2024.00089.

Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014:1–11, 2014. doi: 10.1155/2014/781670.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002. doi: 10.1142/S0218488502001648.

Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE Pervasive Computing*, 16(4):62–74, 2017. doi: 10.1109/MPRV.2017.3971131.

Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R. Schinazi, and Markus Gross. Affective state prediction based on semi-supervised learning from smartphone touch data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*, pp. 1–13. ACM, 2020. doi: 10.1145/3313831.3376504.

Zhiyuan Wang, Mark Rucker, Emma R. Toner, Maria A. Larrazabal, Mehdi Boukhechba, Bethany A. Teachman, and Laura E. Barnes. Understanding privacy risks versus predictive benefits in wearable sensor-based digital phenotyping: A quantitative cost-benefit analysis. In *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, pp. 1–4. IEEE, 2023. doi: 10.1109/BSN58485.2023.10331378.

Gary M. Weiss, Kenichi Yoneda, and Thaier Hayajneh. Smartphone and smartwatch-based biometrics using activities of daily living. *IEEE Access*, 7:133190–133202, 2019. doi: 10.1109/ACCESS.2019.2940729.

Frank Wilcoxon. Individual comparisons by ranking methods, 1945.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole Rashomon set of sparse decision trees. In *Advances in Neural Information Processing Systems 35*, volume 35, pp. 14071–14084. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi: 10.52202/068431-1023.

William J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. doi: 10.1002/1097-0142(1950)3:1⟨32::AID-CNCR2820030106⟩3.0.CO;2-3.

Chudi Zhong, Zhi Chen, Jiachang Liu, Margo Seltzer, and Cynthia Rudin. Exploring and interacting with the set of good sparse generalized additive models. In *Neural Information Processing Systems*, volume 36, pp. 56673–56699. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-2476.

Table 2: The datasets. The *authors'* features shiped with their dataset; we compute *our* features with our own window and channel choices. The last column counts the LOSO folds in which some tree in the approximate target task Rashomon set uses no flagged feature (§5.3).

Dataset	Subjects	Classes	Features	Chance (%) target	Chance (%) identity	Removed	Folds avoiding
ShearSense-Gesture	18	9	33, ours	11.1	5.6	13	4 / 18
WISDM (phone accelerometer)	44	18	46, authors'	5.6	2.3	26	15 / 44
WISDM (phone gyroscope)	50	18	46, authors'	5.6	2.0	13	32 / 50
WISDM (watch accelerometer)	50	18	46, authors'	5.6	2.0	27	0 / 50
WISDM (watch gyroscope)	50	18	46, authors'	5.6	2.0	3	3 / 50
PAMAP2	9	12	124, ours	8.3	11.1	0	9 / 9
OPPORTUNITY (locomotion)	4	4	180, ours	25.0	25.0	4	4 / 4
OPPORTUNITY (both arms)	4	17	180, ours	5.9	25.0	1	4 / 4
ExtraSensory	30	6	145, authors'	16.7	3.3	3	30 / 30
Diabetes 130-US Hospitals	30	2	40, authors'	50.0	3.3	3	30 / 30

Appendix contents

A Datasets . . . 16
B Rashomon sets: construction and checks . . . 20
C Full results . . . 22
D Per-dataset results . . . 27
E The cut rule . . . 28
F Personalization . . . 28
G Differential privacy . . . 29
H Window overlap . . . 29
I Reasons to accept an accuracy cost . . . 29

A Datasets

This appendix describes how each dataset is used in our experiments. For each dataset, we define the subject, target task, retained features, and preprocessing choices. Where we compute features, the window length, step size, and selected channels are part of our preprocessing and do not necessarily match the original benchmark protocol.

ShearSense-Gesture. A soft capacitive sensor recorded normal stress and two-dimensional shear on a grid of 36 taxels at about 43 Hz (McLaren et al., 2024). Each trial contains one of nine affective gestures and is resampled to 86 frames, or about two seconds, following the original study. The public release contains 8,295 trials from 18 participants, including two participants excluded from the original analysis because of sensor degradation. We apply quality-control criteria to individual trials. A trial fails the dropout screen when more than 10% of its frames have a spatial standard deviation below 10^{-3} on the normal channel. It fails the discontinuity screen when any frame-to-frame jump in the summed normal signal exceeds five standard deviations of that trial's jumps. The baselines use the 8,278 trials that pass the dropout screen. The Rashomon-set analyses use the 7,154 trials that pass both screens and have all features defined. We compute 33 features per trial in eight mechanically motivated families (§A.1).

WISDM. Fifty subjects performed 18 activities with a phone in a pocket and a watch on the wrist (Weiss et al., 2019). Each device records an accelerometer and a gyroscope. We window the four device–sensor streams separately and treat each as a separate dataset. The release provides 91 features per ten-second window. We retain the 46 features used in the WISDM papers and exclude the MFCC, cosine, and correlation features, which the dataset documentation describes as experimental. In the phone accelerometer stream, six subjects' files are byte-identical copies of

subject 1610's 404 rows and also assign those rows to subject 1610. We remove those six files, leaving 44 subjects in that dataset.

PAMAP2. Nine subjects wore three inertial measurement units on the hand, chest, and ankle, together with a heart-rate monitor (Reiss, 2012). We use the protocol sessions and exclude the transient activity label as recommended in the dataset documentation. For each inertial unit, we use the ± 16 g accelerometer, gyroscope, and magnetometer. We exclude the ± 6 g accelerometer, which saturates, and the orientation channels, which the documentation marks as invalid. Heart rate is sampled more slowly than the motion channels, so we carry each reading forward for at most one second. We form 512-sample windows, corresponding to 5.12 seconds, with a one-second step. We retain a window only when all of its samples have the same activity label. For each channel, we compute the mean, standard deviation, minimum, and maximum, giving 124 features (§A.1). One subject's protocol file contains only the rope-jumping activity.

OPPORTUNITY. Four subjects performed daily activities and a drill session in a sensor-rich environment (Roggen et al., 2010). We use the five body-worn inertial units on the back and upper and lower arms, with acceleration, angular rate, and magnetic field. We exclude the quaternion channels and shoe units. We pool the runs for each subject and form two-second windows with a one-second step at 30 Hz without crossing run boundaries. We retain a window only when all samples have the same non-null label. For each channel, we compute the mean, standard deviation, minimum, and maximum, giving 180 features (§A.1). The two label tracks define two datasets, locomotion with 4 classes and mid-level gestures of both arms with 17 classes.

ExtraSensory. Sixty users carried their own phones during daily activities and labeled their context at one-minute intervals (Vaizman et al., 2017). Because the labels are multi-label, we use six primary activities: lying down, sitting, walking, running, bicycling, and standing. We retain only minutes with exactly one of these six labels. We retain six phone sensor families, giving 145 features, and exclude five other families. The excluded families are missing for entire users because most users did not wear the watch and many did not enable location services. Encoding this missingness as zeros would make missingness associated with user identity. Removing all incomplete minutes would exclude 13 of 30 users. We therefore retain only complete minutes so that all phases analyze the same rows. Among the 55 users with at least 200 complete single-activity minutes, we sample 30 users at random. We retain at most 500 minutes per user, sampled within user so that each user's observed activity distribution is preserved.

Diabetes 130-US Hospitals. The release contains 101,766 inpatient encounters from 71,518 patients at 130 US hospitals between 1999 and 2008 (Clore et al., 2014; Strack et al., 2014). The original study retained one encounter per patient to obtain independent observations. Our identity task requires repeated encounters, so we retain them. The target task is readmission within 30 days versus all other outcomes, matching the binary target used in the original study. We sample 30 patients at random from the 39 patients with at least 15 encounters. This sampling produces a cohort with frequent readmission, in which "not readmitted" decreases from 54% of all encounters to about 2%. The reported results therefore describe this selected cohort rather than the full hospital population. The 40 features fall into seven families: demographics, admission, prior use of care, workload, laboratory results, diagnoses by chapter group, and medications. Race and sex are constant within a patient and can therefore contribute to identity prediction. We exclude weight, which is 97% missing, payer code, medical specialty, drugs prescribed in fewer than 1% of encounters, and both identifiers.

A.1 Features we compute

Three of the ten datasets did not release features, so we compute them from the touch tensor of each trial for ShearSense-Gesture and from windows of the sensor streams for PAMAP2 and OPPORTUNITY. The other seven datasets use the features their authors released, as described above.

ShearSense-Gesture: why these features. In the original ShearSense study, McLaren et al. (2024) classified gestures using five summary features (mean normal stress, mean shear magnitude, contact area, peak frequency, shear-to-normal ratio). The feature families are motivated by

touch biomechanics literature. Hauser et al. (2019) documented that force magnitude, velocity, and contact area vary substantially across individuals even when performing the same communicative intent, suggesting that features capturing *how* someone applies force (spatial distribution, temporal modulation) carry different information from features capturing *how much* force is applied. Cang et al. (2015) showed that dynamic surface features (pressure temporal profiles, contact-area changes) capture both gesture and individual differences in affective touch, and Croy et al. (2016) found that stroking velocity profiles are individually tuned to CT-afferent-optimal ranges, motivating a feature that directly measures the proportion of touch within that range. McLaren et al. (2024) noted strong individual variation across their five metrics (Table 6 in the original paper) but had no features for the spatial structure of the shear field (curl, divergence, coherence), centroid trajectory dynamics, or rhythmic profile, all of which plausibly differ between gesture types and between individuals for mechanically distinct reasons. We engineered 33 scalar features per trial, in eight mechanically motivated families (Table 3), to decompose the touch signal with sufficient granularity to test overlap of gesture- and identity-discriminative features.

Table 3: The 33 ShearSense-Gesture features, grouped by family. Features marked † are new relative to McLaren et al. (2024); features marked ‡ are retained at the cut (§5.5).

Family & summary	**Features (count)**
Pressure: normal force magnitude and spatial extent	(5) mean_normal‡, std_normal, max_normal‡, contact_area_mean, contact_area_std‡
Shear scalar: tangential loading magnitude, uniformity, and shear-to-normal ratio	(5) mean_shear†, std_shear†‡, max_shear†, shear_to_normal‡, shear_uniformity†
Temporal dynamics: frequency content, temporal shape, self-similarity of the normal-force trace	(5) peak_freq, spectral_entropy†‡, temporal_centroid†, autocorr_lag1†, rise_time†
Spatial / centroid†: centroid position, motion, and path length	(6) centroid_x_mean, centroid_y_mean, centroid_x_std‡, centroid_y_std, centroid_path_length‡, centroid_velocity_mean
Shear vector field†: curl, divergence, coherence, angular dispersion of the shear field	(4) shear_curl_mean, shear_divergence_mean, shear_coherence, shear_angle_std‡
Contact dynamics†: contact-area change, onset count, cumulative pressure	(3) contact_area_delta, contact_onset_count‡, pressure_auc‡
Periodicity†: rhythmic regularity, duty cycle, spectral flux	(3) inter_peak_cv‡, duty_cycle‡, spectral_flux
Sliding / C-tactile†: sliding fraction and proportion of centroid velocities in the CT range (1–10 cm/s)	(2) sliding_ratio‡, ct_velocity_proportion
Total	**33**

ShearSense-Gesture: how the features are computed. Let $a_0 \in \mathbb{R}^{F\times H\times W}$ denote the normal-force array for a trial of $F = 86$ frames on an $H \times W = 11 \times 11$ grid, and $a_1, a_2 \in \mathbb{R}^{F\times H\times W}$ the two shear channels. We compute all means, standard deviations, and maxima over *all* taxels and frames, including those outside the contact region, unless otherwise specified. The contact mask at frame t is $M_t = \{(i, j) : a_{0,tij} > 0.05 \max a_0\}$, with $\max a_0$ the trial's global maximum, and $f_s = 43$ Hz is the frame rate.

Pressure.

$$\begin{aligned}
\text{mean_normal} &= \frac{1}{FHW}\sum_{t=1}^{F}\sum_{i=1}^{H}\sum_{j=1}^{W} a_{0,tij},\\
\text{std_normal} &= \operatorname{std}\left(\{a_{0,tij}\}_{t,i,j}\right),\\
\text{max_normal} &= \max_{t,i,j} a_{0,tij},\\
\text{contact_area_mean} &= \frac{1}{F}\sum_{t=1}^{F}|M_t|,\\
\text{contact_area_std} &= \operatorname{std}_t|M_t|.
\end{aligned}$$

Shear scalar. With $\|\mathbf{s}_{t,ij}\|_2 = \sqrt{a_{1,tij}^2 + a_{2,tij}^2}$ and $\bar{s}_{ij} = \frac{1}{F}\sum_{t=1}^{F} \|\mathbf{s}_{t,ij}\|_2$,

$$\begin{aligned}
\text{mean_shear} &= \frac{1}{FHW}\sum_{t,i,j} \|\mathbf{s}_{t,ij}\|_2, \\
\text{std_shear} &= \operatorname{std}\left(\{\|\mathbf{s}_{t,ij}\|_2\}_{t,i,j}\right), \\
\text{max_shear} &= \max_{t,i,j} \|\mathbf{s}_{t,ij}\|_2, \\
\text{shear_to_normal} &= \frac{\text{mean_shear}}{\text{mean_normal}}, \\
\text{shear_uniformity} &= \frac{1}{1 + \mathrm{CV}(\bar{s}_{ij})},
\end{aligned}$$

where CV is the coefficient of variation across taxels.

Temporal dynamics. Let $P(t) = \sum_{i,j}\left(a_{0,tij} + \|\mathbf{s}_{t,ij}\|_2\right)$ and let $S(f_k)$ be the power spectrum of $P - \bar{P}$.

$$\begin{aligned}
\text{peak_freq} &= \arg\max_{f_k \le 5\,\text{Hz}} S(f_k), \\
\text{spectral_entropy} &= -\sum_{f_k \le 5\,\text{Hz}} \hat{S}(f_k)\log_2 \hat{S}(f_k), \quad \hat{S} = S \Big/ \sum_{f_k \le 5\,\text{Hz}} S, \\
\text{temporal_centroid} &= \frac{\sum_t t\, P(t)^2}{f_s \sum_t P(t)^2}, \\
\text{autocorr_lag1} &= \frac{\sum_{t=1}^{F-1}(P(t) - \bar{P})(P(t+1) - \bar{P})}{\sum_t (P(t) - \bar{P})^2}, \\
\text{rise_time} &= \frac{t_{90} - t_{10}}{f_s},
\end{aligned}$$

where t_{10} and t_{90} are the first frames where $P(t)$ exceeds 10% and 90% of its maximum.

Spatial and centroid. The centroid at frame t is $c_x(t) = \sum_{i,j} j\, a_{0,tij} / \sum_{i,j} a_{0,tij}$ and $c_y(t) = \sum_{i,j} i\, a_{0,tij} / \sum_{i,j} a_{0,tij}$, with $\Delta\mathbf{c}(t) = (c_x(t+1) - c_x(t),\, c_y(t+1) - c_y(t))$. The features are $\operatorname{mean}_t c_x(t)$, $\operatorname{mean}_t c_y(t)$, $\operatorname{std}_t c_x(t)$ and $\operatorname{std}_t c_y(t)$, the path length $\sum_{t=1}^{F-1} \|\Delta\mathbf{c}(t)\|$, and the mean velocity $f_s \operatorname{mean}_t \|\Delta\mathbf{c}(t)\|$.

Shear vector field. All four features are averaged over frames with at least 3 active taxels.

$$\begin{aligned}
\text{shear_curl_mean} &= \operatorname{mean}_t \operatorname{mean}_{i,j} \left(\frac{\partial a_2}{\partial x} - \frac{\partial a_1}{\partial y}\right)_{t,ij}, \\
\text{shear_divergence_mean} &= \operatorname{mean}_t \operatorname{mean}_{i,j} \left(\frac{\partial a_1}{\partial x} + \frac{\partial a_2}{\partial y}\right)_{t,ij}, \\
\text{shear_coherence} &= \frac{\sqrt{(\operatorname{mean} a_1)^2 + (\operatorname{mean} a_2)^2}}{\operatorname{mean} \|\mathbf{s}\|_2}, \\
\text{shear_angle_std} &= \sqrt{-2\log R}, \quad R = \sqrt{(\operatorname{mean}\cos\phi)^2 + (\operatorname{mean}\sin\phi)^2},
\end{aligned}$$

where $\phi = \operatorname{arctan2}(a_2, a_1)$.

Contact dynamics.

$$\begin{aligned}
\text{contact_area_delta} &= \frac{1}{F-1}\sum_{t=1}^{F-1} \big|\, |M_{t+1}| - |M_t| \,\big|, \\
\text{contact_onset_count} &= \sum_{t=2}^{F} \mathbb{I}\left[|M_{t-1}| = 0 \wedge |M_t| > 0\right], \\
\text{pressure_auc} &= \frac{1}{f_s}\sum_{t,i,j} a_{0,tij}.
\end{aligned}$$

Periodicity.

$$\text{inter_peak_cv} = \frac{\text{std}(\Delta t_k)}{\text{mean}(\Delta t_k)},$$

$$\text{duty_cycle} = \frac{1}{F}\sum_{t=1}^{F} \mathbb{I}\left[|M_t| > 0\right],$$

$$\text{spectral_flux} = \text{mean}_n \left\|S_{n+1} - S_n\right\|_1,$$

where Δt_k are the intervals between peaks of $P(t)$ that reach 10% of its maximum and lie at least 3 frames apart, and S_n is the magnitude spectrum of P over the n-th 16-frame window, taken with an 8-frame hop.

Sliding and C-tactile. Let $v_c(t) = \|\Delta \mathbf{c}(t)\| \cdot p \cdot f_s$, with grid pitch $p = 0.2\,\text{cm}$ (2.0 mm), and let $\bar{s}(t)$ be the mean shear magnitude at frame t.

$$\text{sliding_ratio} = \frac{1}{F-1}\sum_{t=1}^{F-1} \mathbb{I}\left[\|\Delta \mathbf{c}(t)\| > 0.5 \wedge \bar{s}(t) > 0.1 \cdot \text{mean_shear}\right],$$

$$\text{ct_velocity_proportion} = \frac{1}{F-1}\sum_{t=1}^{F-1} \mathbb{I}\left[1 \le v_c(t) \le 10\right],$$

with v_c in cm/s. The CT-afferent velocity range (1–10 cm/s) follows Morrison et al. (2010) and Croy et al. (2016).

PAMAP2 and OPPORTUNITY: window statistics. Both datasets release continuous sensor streams without precomputed features, so we cut each stream into windows and summarize every channel within a window by its mean, standard deviation, minimum and maximum, following the segmentation and feature-extraction steps of the activity recognition chain for body-worn inertial sensors (Bulling et al., 2014). App. A gives the window length, the step and the channels for each dataset. Each feature is therefore one statistic of one sensor channel, so a flagged feature identifies the channel and the statistic that carry identity. The window length, the step and the four statistics are our choices and do not reproduce the benchmark protocols of the dataset papers. No consensus exists on which window size to use in activity recognition (Banos et al., 2014). We drop a window whose samples carry more than one label, instead of assigning it the majority label, so no row mixes two activities, and no OPPORTUNITY window crosses the boundary between two runs. The step is shorter than the window, so consecutive windows overlap. In App. H, we report tests on whether the overlap raises identity accuracy.

B Rashomon sets: construction and checks

The main analysis uses sampled Rashomon sets from two model classes. We compare them with sets sampled from the grid alone and with tree sets that TREEFARMS and PRAXIS enumerate. Relative to the grid alone, the varied candidates remove the flags that only the tree class produces and some GAM flags near the threshold.

Construction. The tree sample is drawn from a grid over depth $\{2, 3, 4\}$, minimum samples per split $\{2, 5, 10, 20\}$, minimum impurity decrease $\{0, 0.001, 0.005, 0.01\}$, and features per split $\{p, \sqrt{p}, \log_2 p\}$. The grid contains 144 configurations, each fitted with ten seeds, for 1,440 trees in total, using Gini impurity. We add 4,000 trees that draw at random the depth from $\{2, 3, 4\}$, minimum samples per split from $\{2, 5, 10, 20\}$, minimum impurity decrease from $\{0, 0.001, 0.005, 0.01\}$, features per split from $\{p, \sqrt{p}, \log_2 p, 0.1p, 0.3p, 0.5p, 0.7p\}$, minimum leaf size from $\{1, 5, 20\}$, best or random split thresholds, and the seed, and half of these trees see only a random half of the features. The GAM sample fits EBMs without interaction terms (Nori et al., 2019) at 24 distinct boosting-round budgets spaced geometrically between 1 and 5,000. We add 200 EBMs that draw at random the number of boosting rounds from $\{100, 300, 1000, 5000\}$, the number of bins from $\{16, 32, 64, 256\}$, the learning rate from $\{0.005, 0.015, 0.05\}$, the minimum leaf size from $\{2, 4, 16\}$, the number of outer bags from $\{4, 14\}$, and the seed. A model enters the Rashomon set when its held-out loss is at most the best held-out loss among all candidates of its class plus

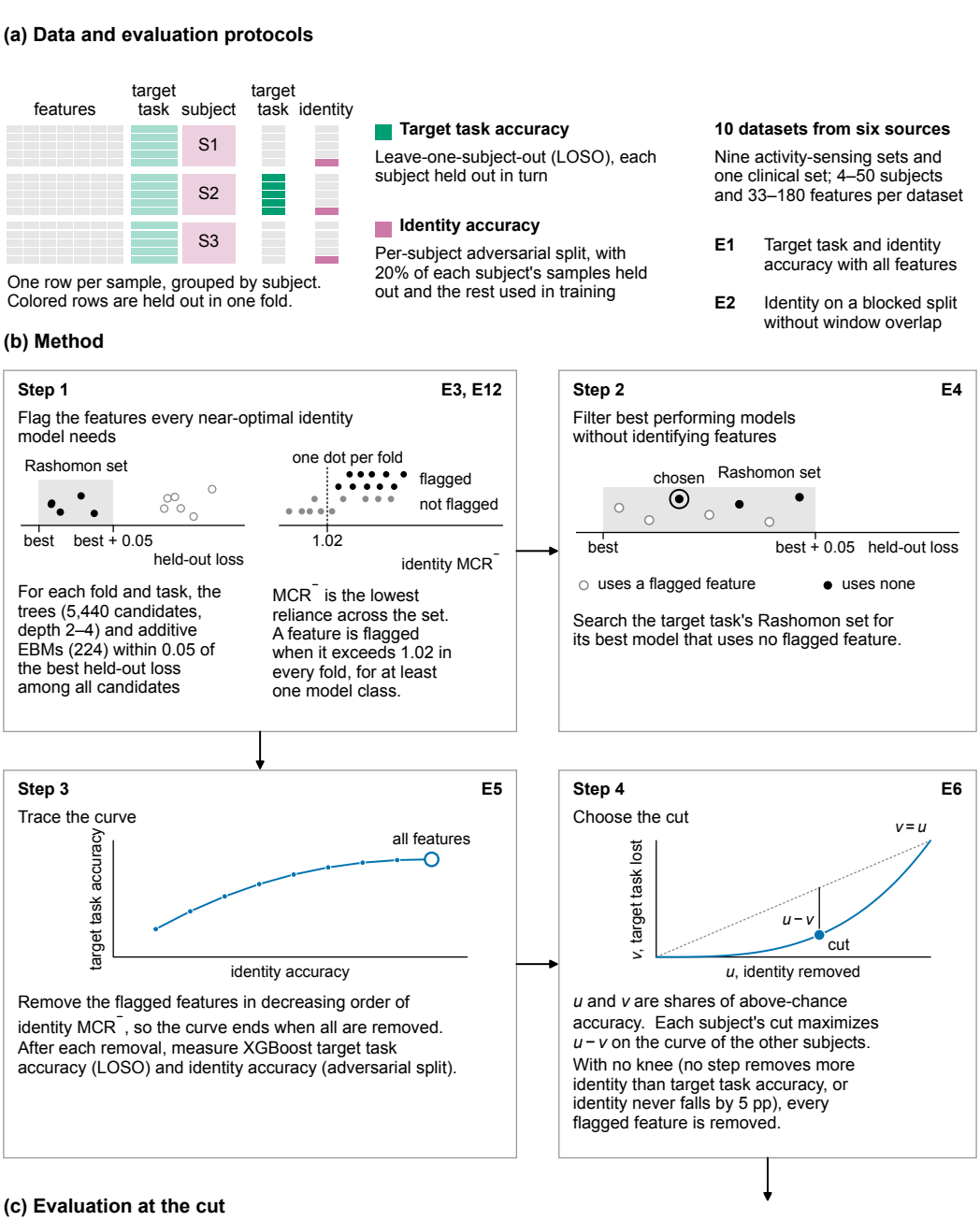


Figure 4: Every analysis step, from the data to the tests. (a) Each sample carries a target task label and a subject. Target task accuracy is scored by leave-one-subject-out (LOSO) and identity accuracy by the per-subject adversarial split, which holds out 20% of each subject's samples. (b) shows the method and the experiments that test each step (E3 to E6 and E12). Step 1 feeds both Step 2 and Step 3. MCR scores every member of a fold on the same 50 permutations per feature. (c) lists the experiments run at the cut (E7 to E11). The curves in (b) are schematic.

$\varepsilon = 0.05$. For identity on ShearSense, this leaves 164 GAMs and 275 trees per fold on average. The ShearSense target task sets contain between 36 and 682 trees per fold. We estimate model reliance with 50 permutations per feature and set it to 1 when a model's loss on the intact data is zero. Every model in a fold's set is scored on the same permutations.

Exact tree sets. We compare our tree sets with tree sets that TREEFARMS (Xin et al., 2022) and PRAXIS (Heile et al., 2026a) enumerate. On the clinical dataset, the two methods return the same trees at depths 2 and 3. We compute identity MCR^- from PRAXIS sets of depth-3 trees on ShearSense, WISDM watch accelerometer and the clinical dataset, with each feature binarized at 24 thresholds taken from a gradient-boosted model. No feature reaches $\mathrm{MCR}^- > 1.02$ in these sets, which agrees with our sampled tree sets on the same datasets.

Sets sampled from the grid alone. For comparison, we also sample Rashomon sets from the grid alone, without the varied candidates, keeping the fits within ε of the best grid candidate. Table 4 counts, for each dataset, the features whose mean identity MCR^- over folds exceeds 1.02 under the grid-only and our sets, and the features our sets flag under the per-fold rule of §4 at three thresholds. On 10 of the 10 datasets, every feature our sets flag is also flagged by the grid-only sets.

Table 4: Flagged features under the grid-only sets and ours. The middle columns apply a looser rule, mean identity MCR^- over folds above 1.02 in either model class. The last three columns apply the per-fold rule, $\mathrm{MCR}^- > \tau$ in every fold for some model class, to our sets.

		Mean over folds > 1.02		Every fold $> \tau$		
Dataset	Features	Grid only	Ours	$\tau = 1.01$	1.02	1.03
ShearSense-Gesture	33	19	14	15	13	12
WISDM (phone accelerometer)	46	30	28	27	26	24
WISDM (phone gyroscope)	46	18	14	20	13	8
WISDM (watch accelerometer)	46	30	30	33	28	26
WISDM (watch gyroscope)	46	4	3	6	3	3
PAMAP2	124	21	8	0	0	0
OPPORTUNITY (locomotion)	180	14	7	4	4	3
OPPORTUNITY (both arms)	180	11	3	1	1	1
ExtraSensory	145	6	3	3	3	3
Diabetes 130-US Hospitals	40	15	6	3	3	3

Why most features get flagged by the GAM set. Under the looser rule, mean identity MCR^- over folds above 1.02, the grid-only tree sets flag 32 features, all on PAMAP2, the two OPPORTUNITY tracks and the clinical dataset, and our tree sets flag 10. Under the per-fold rule, our tree sets flag one feature, on OPPORTUNITY locomotion. A tree of depth at most 4 makes at most 15 splits, so it uses few features, and many small feature subsets reach near-optimal identity accuracy. Because half of the varied tree candidates see a random half of the features, some near-optimal tree leaves out almost any given feature, and tree identity MCR^- is at most 1 in 93% to 100% of fold-feature cells on every dataset. A GAM adds one term per feature, so its identity accuracy builds up across many features. To test whether the GAM flags come from a narrow candidate set, we added 200 EBMs that each see a random half of the 33 features to the ShearSense identity candidates. The best of them reaches identity accuracy 9.6 ± 0.5 pp below the best candidate, mean $\pm$ SD over five folds and never closer than 8.9 pp, so none enters the set at $\varepsilon = 0.05$ and the same 13 features are flagged with identical per-fold MCR^-. A flag therefore describes the sampled sets, and we ran this check on ShearSense only.

C Full results

These tables report the comparisons used in §5. For each comparison, they report both arms as mean $\pm$ SD across subjects, the change in percentage points, the effect size r, and the Holm-adjusted p value for the stated family. Arm A is the reference and arm B is the comparison, so a negative change means that B is lower. A dagger marks comparisons without a statistical test because the arms are identical for every subject or fewer than two subjects remain. For datasets with up to 50 subjects, the Wilcoxon test uses the exact null distribution. When differences are tied or zero, it uses the

exact conditional distribution of the mid-rank statistic. For datasets with more than 50 subjects, it uses the normal approximation. OPPORTUNITY has only 4 subjects, for which a two-sided signed-rank test cannot reach $p < 0.05$, so we report those results descriptively. PAMAP2's 9 subjects allow $p = 0.004$ at best, which Holm correction raises to 0.055 among the comparisons at the cut and to 0.168 among the selection-rule comparisons, so PAMAP2's results in those two tables are also descriptive, and the selection-rule counts leave it out. The column r gives each comparison's matched-pairs rank-biserial correlation as its effect size, signed like the change. The text and tables are generated from the same digest.

The cut (E4, E7). Arm A uses all features and arm B uses the features retained after the cut. The target task uses LOSO, and identity uses the adversarial split. The tree row compares the best Rashomon tree using all features with the best tree that avoids the removed features. Holm correction is applied across the 37 comparisons in this table.

Table 5: Accuracy with all features (A) and with the features retained after the cut (B), mean ± SD across subjects (E4, E7). The target task uses LOSO and identity the adversarial split. Δ is B − A in pp, and p is Holm-adjusted across the 37 comparisons in this table.

Dataset	Comparison	n	A (%)	B (%)	Δ (pp)	r	p_{Holm}
ShearSense-Gesture	XGB target	18	55.7 ± 14.5	51.3 ± 13.8	-4.4	-0.86	0.010
	XGB identity	18	65.6 ± 10.1	41.7 ± 14.2	-24.0	-1.00	< 0.001
	Rashomon tree target	18	31.5 ± 11.2	30.7 ± 10.5	-0.7	-0.11	1.000
	XGB target (nested cut)	18	55.7 ± 14.5	51.3 ± 13.8	-4.4	-0.86	0.010
	XGB identity (nested cut)	18	65.6 ± 10.1	41.7 ± 14.2	-24.0	-1.00	< 0.001
WISDM (phone accelerometer)	XGB target	44	30.8 ± 8.1	28.8 ± 6.5	-2.0	-0.36	0.426
	XGB identity	44	90.5 ± 6.0	32.5 ± 12.8	-58.0	-1.00	< 0.001
	Rashomon tree target	44	22.2 ± 7.3	20.8 ± 6.4	-1.5	-0.23	1.000
	XGB target (nested cut)	44	30.8 ± 8.1	28.7 ± 6.6	-2.1	-0.38	0.351
	XGB identity (nested cut)	44	90.4 ± 5.9	32.9 ± 13.2	-57.5	-1.00	< 0.001
WISDM (phone gyroscope)	XGB target	50	31.6 ± 6.6	27.4 ± 7.2	-4.2	-0.95	< 0.001
	XGB identity	50	51.1 ± 14.9	35.7 ± 13.8	-15.4	-1.00	< 0.001
	Rashomon tree target	50	20.4 ± 6.3	18.1 ± 5.9	-2.3	-0.53	0.012
	XGB target (nested cut)	50	31.6 ± 6.6	27.4 ± 7.2	-4.2	-0.95	< 0.001
	XGB identity (nested cut)	50	51.1 ± 14.7	35.4 ± 13.8	-15.6	-1.00	< 0.001
WISDM (watch accelerometer)	XGB target	50	66.5 ± 15.9	52.4 ± 15.3	-14.1	-1.00	< 0.001
	XGB identity	50	63.0 ± 16.4	24.0 ± 16.5	-39.0	-1.00	< 0.001
	Rashomon tree target	50	33.8 ± 10.6	27.8 ± 8.5	-6.0	-0.86	< 0.001
	XGB target (nested cut)	50	66.5 ± 15.9	52.3 ± 15.3	-14.1	-1.00	< 0.001
	XGB identity (nested cut)	50	63.4 ± 16.1	24.3 ± 16.4	-39.1	-1.00	< 0.001
WISDM (watch gyroscope)	XGB target	50	58.5 ± 15.1	53.3 ± 15.7	-5.2	-0.98	< 0.001
	XGB identity	50	36.6 ± 12.7	27.9 ± 11.1	-8.7	-0.98	< 0.001
	Rashomon tree target	50	38.4 ± 12.0	34.6 ± 11.3	-3.8	-0.72	< 0.001
	XGB target (nested cut)	50	58.5 ± 15.1	53.3 ± 15.7	-5.2	-0.98	< 0.001
	XGB identity (nested cut)	50	36.8 ± 12.8	28.0 ± 11.1	-8.8	-0.97	< 0.001
PAMAP2	XGB target	9	87.0 ± 9.0	87.0 ± 9.0	0.0	–	–[†]
	XGB identity	9	100.0 ± 0.1	100.0 ± 0.1	0.0	–	–[†]
	Rashomon tree target	9	57.7 ± 22.5	57.7 ± 22.5	0.0	–	–[†]
OPPORTUNITY (locomotion)	XGB target	4	93.7 ± 4.8	93.8 ± 4.6	$+0.1$	0.20	1.000
	XGB identity	4	99.0 ± 0.6	98.9 ± 0.6	-0.1	-0.40	1.000
	Rashomon tree target	4	86.0 ± 2.7	86.0 ± 2.7	0.0	–	–[†]
OPPORTUNITY (both arms)	XGB target	4	82.7 ± 4.7	82.9 ± 4.6	$+0.2$	0.80	1.000
	XGB identity	4	99.0 ± 0.9	98.6 ± 0.7	-0.4	-0.67	1.000
	Rashomon tree target	4	53.3 ± 9.0	53.3 ± 9.0	0.0	–	–[†]
ExtraSensory	XGB target	30	70.7 ± 9.1	71.1 ± 9.1	$+0.5$	0.25	1.000
	XGB identity	30	98.3 ± 2.5	95.0 ± 5.6	-3.3	-0.98	0.002
	Rashomon tree target	30	69.4 ± 9.2	69.5 ± 9.2	$+0.1$	0.33	1.000
Diabetes 130-US Hospitals	XGB target	30	54.1 ± 13.1	53.7 ± 13.8	-0.4	-0.21	1.000
	XGB identity	30	93.9 ± 12.6	72.5 ± 28.5	-21.4	-0.91	0.010
	Rashomon tree target	30	56.7 ± 10.3	57.7 ± 11.2	$+1.0$	0.10	1.000
	XGB target (nested cut)	28	54.8 ± 13.3	52.6 ± 13.5	-2.2	-0.42	0.922
	XGB identity (nested cut)	28	93.5 ± 12.9	72.9 ± 28.0	-20.5	-0.90	0.017

Table 6: Equivalence of the target task change at the cut (E7), for the arms of Table 5. The TOST asks whether B − A lies within ±5 pp, and p is the larger of its two one-sided Wilcoxon signed-rank tests, Holm-adjusted across the 9 datasets. A p below 0.05 means that the change lies within the margin.

Dataset	Comparison	n	A (%)	B (%)	Δ (pp)	r	p_{Holm}
ShearSense-Gesture	XGB target within margin	18	55.7 ± 14.5	51.3 ± 13.8	-4.4	–	0.826

Dataset	Comparison	n	A (%)	B (%)	Δ (pp)	r	p_{Holm}
WISDM (phone accelerometer)	XGB target within margin	44	30.8 ± 8.1	28.8 ± 6.5	-2.0	–	0.004
WISDM (phone gyroscope)	XGB target within margin	50	31.6 ± 6.6	27.4 ± 7.2	-4.2	–	0.129
WISDM (watch accelerometer)	XGB target within margin	50	66.5 ± 15.9	52.4 ± 15.3	-14.1	–	1.000
WISDM (watch gyroscope)	XGB target within margin	50	58.5 ± 15.1	53.3 ± 15.7	-5.2	–	1.000
PAMAP2	XGB target within margin	9	87.0 ± 9.0	87.0 ± 9.0	0.0	–	–[†]
OPPORTUNITY (locomotion)	XGB target within margin	4	93.7 ± 4.8	93.8 ± 4.6	$+0.1$	–	0.312
OPPORTUNITY (both arms)	XGB target within margin	4	82.7 ± 4.7	82.9 ± 4.6	$+0.2$	–	0.312
ExtraSensory	XGB target within margin	30	70.7 ± 9.1	71.1 ± 9.1	$+0.5$	–	< 0.001
Diabetes 130-US Hospitals	XGB target within margin	30	54.1 ± 13.1	53.7 ± 13.8	-0.4	–	0.184

Deep model (E9). Arms A and B use all features and the retained features, respectively, with TabPFN v2. Holm correction is applied across the 18 comparisons in this table.

Table 7: TabPFN v2 accuracy with all features (A) and with the retained features (B), mean $\pm$ SD across subjects (E9). Δ is B $-$ A in pp, and p is Holm-adjusted across the 18 comparisons in this table.

Dataset	Comparison	n	A (%)	B (%)	Δ (pp)	r	p_{Holm}
ShearSense-Gesture	TabPFN target	18	59.5 ± 14.8	55.1 ± 14.8	-4.3	-0.96	0.001
	TabPFN identity	18	73.4 ± 9.0	48.4 ± 15.0	-24.9	-1.00	< 0.001
WISDM (phone accelerometer)	TabPFN target	44	30.5 ± 9.1	29.9 ± 6.4	-0.5	0.10	1.000
	TabPFN identity	44	91.4 ± 6.3	29.0 ± 17.5	-62.4	-1.00	< 0.001
WISDM (phone gyroscope)	TabPFN target	50	33.0 ± 6.8	29.3 ± 8.3	-3.7	-0.67	< 0.001
	TabPFN identity	50	50.1 ± 19.2	30.3 ± 16.9	-19.8	-1.00	< 0.001
WISDM (watch accelerometer)	TabPFN target	50	67.0 ± 15.9	54.1 ± 15.7	-12.8	-1.00	< 0.001
	TabPFN identity	50	58.6 ± 22.3	15.6 ± 21.3	-43.0	-1.00	< 0.001
WISDM (watch gyroscope)	TabPFN target	50	61.3 ± 15.8	56.4 ± 15.7	-4.9	-0.96	< 0.001
	TabPFN identity	50	27.2 ± 19.0	17.6 ± 16.9	-9.5	-0.92	< 0.001
PAMAP2	TabPFN target	9	77.0 ± 12.4	77.0 ± 12.4	0.0	–	–[†]
	TabPFN identity	9	100.0 ± 0.0	100.0 ± 0.0	0.0	–	–[†]
OPPORTUNITY (locomotion)	TabPFN target	4	93.7 ± 4.9	93.8 ± 4.7	$+0.1$	0.40	1.000
	TabPFN identity	4	99.7 ± 0.3	99.4 ± 0.9	-0.4	-1.00	1.000
OPPORTUNITY (both arms)	TabPFN target	4	84.7 ± 6.3	86.1 ± 4.8	$+1.4$	1.00	0.875
	TabPFN identity	4	99.4 ± 0.8	99.5 ± 0.4	$+0.1$	0.33	1.000
ExtraSensory	TabPFN target	30	69.9 ± 9.8	70.5 ± 9.5	$+0.6$	0.06	1.000
	TabPFN identity	30	98.6 ± 2.4	94.7 ± 6.1	-3.8	-0.98	< 0.001
Diabetes 130-US Hospitals	TabPFN target	30	50.0 ± 13.4	49.3 ± 14.2	-0.6	-0.13	1.000
	TabPFN identity	30	94.7 ± 14.9	75.6 ± 26.7	-19.2	-1.00	0.001

Personalization (E8). Arm A includes none of the test person's data in training and arm B includes all available training data from that person. Results are averaged over four models, random forest, extra trees, XGBoost (Chen & Guestrin, 2016), and a GAM, and five seeds per person. We report this comparison for both the full and retained feature sets. Holm correction is applied across the 20 comparisons in this table.

Table 8: Target task accuracy with none (A) and all (B) of the test person's data in training, on all features and on the retained features, mean $\pm$ SD across subjects (E8). Δ is B $-$ A in pp, and p is Holm-adjusted across the 20 comparisons in this table.

Dataset	Comparison	n	A (%)	B (%)	Δ (pp)	r	p_{Holm}
ShearSense-Gesture	E all features	18	53.8 ± 14.7	69.3 ± 11.2	$+15.6$	1.00	< 0.001
	E privacy-safe	18	49.2 ± 14.3	60.0 ± 12.0	$+10.8$	1.00	< 0.001
WISDM (phone accelerometer)	E all features	44	28.9 ± 7.8	77.4 ± 7.2	$+48.6$	1.00	< 0.001
	E privacy-safe	44	27.8 ± 6.5	40.0 ± 7.6	$+12.3$	1.00	< 0.001
WISDM (phone gyroscope)	E all features	50	29.8 ± 7.0	49.6 ± 7.4	$+19.7$	1.00	< 0.001
	E privacy-safe	50	25.9 ± 7.3	39.4 ± 7.9	$+13.5$	1.00	< 0.001
WISDM (watch accelerometer)	E all features	50	66.3 ± 16.7	82.1 ± 7.0	$+15.8$	1.00	< 0.001
	E privacy-safe	50	50.6 ± 15.8	57.9 ± 13.6	$+7.3$	1.00	< 0.001
WISDM (watch gyroscope)	E all features	50	57.3 ± 16.2	67.4 ± 10.3	$+10.1$	1.00	< 0.001
	E privacy-safe	50	52.3 ± 16.1	60.8 ± 11.9	$+8.5$	1.00	< 0.001
PAMAP2	E all features	9	89.0 ± 6.4	99.5 ± 0.4	$+10.5$	1.00	0.031
	E privacy-safe	9	89.0 ± 6.4	99.5 ± 0.4	$+10.5$	1.00	0.031
OPPORTUNITY (locomotion)	E all features	4	92.4 ± 4.9	97.9 ± 0.6	$+5.5$	1.00	0.500
	E privacy-safe	4	92.6 ± 4.4	97.9 ± 0.6	$+5.3$	1.00	0.500
OPPORTUNITY (both arms)	E all features	4	83.3 ± 7.4	95.2 ± 0.7	$+11.9$	1.00	0.500
	E privacy-safe	4	83.4 ± 7.4	95.1 ± 0.6	$+11.7$	1.00	0.500
ExtraSensory	E all features	30	71.1 ± 8.7	81.6 ± 6.7	$+10.6$	1.00	< 0.001
	E privacy-safe	30	71.5 ± 8.4	81.1 ± 6.8	$+9.6$	1.00	< 0.001
Diabetes 130-US Hospitals	E all features	30	53.8 ± 15.0	61.8 ± 13.7	$+8.0$	0.55	0.043

Dataset	Comparison	n	A (%)	B (%)	Δ (pp)	r	p_{Holm}
	E privacy-safe	30	53.9 ± 16.7	61.3 ± 13.7	$+7.4$	0.58	0.043

Personalized subset against all features (E8). Arm A uses all features with none of the test person's data in training, and arm B uses the retained features with all available training data from that person. Results are averaged over the same four models and five seeds per person. Holm correction is applied across the 10 comparisons in this table.

Table 9: Target task accuracy of the population model on all features (A) and of the retained features trained with the person's own data (B), mean ± SD across subjects (E8). Δ is B − A in pp, and p is Holm-adjusted across the 10 comparisons in this table.

Dataset	Comparison	n	A (%)	B (%)	Δ (pp)	r	p_{Holm}
ShearSense-Gesture	subset own vs all target	18	53.8 ± 14.7	60.0 ± 12.0	$+6.2$	0.77	0.013
WISDM (phone accelerometer)	subset own vs all target	44	28.9 ± 7.8	40.0 ± 7.6	$+11.2$	1.00	< 0.001
WISDM (phone gyroscope)	subset own vs all target	50	29.8 ± 7.0	39.4 ± 7.9	$+9.6$	0.95	< 0.001
WISDM (watch accelerometer)	subset own vs all target	50	66.3 ± 16.7	57.9 ± 13.6	-8.5	-0.73	< 0.001
WISDM (watch gyroscope)	subset own vs all target	50	57.3 ± 16.2	60.8 ± 11.9	$+3.5$	0.51	0.007
PAMAP2	subset own vs all target	9	89.0 ± 6.4	99.5 ± 0.4	$+10.5$	1.00	0.013
OPPORTUNITY (locomotion)	subset own vs all target	4	92.4 ± 4.9	97.9 ± 0.6	$+5.5$	1.00	0.250
OPPORTUNITY (both arms)	subset own vs all target	4	83.3 ± 7.4	95.1 ± 0.6	$+11.8$	1.00	0.250
ExtraSensory	subset own vs all target	30	71.1 ± 8.7	81.1 ± 6.8	$+10.1$	1.00	< 0.001
Diabetes 130-US Hospitals	subset own vs all target	30	53.8 ± 15.0	61.3 ± 13.7	$+7.5$	0.62	0.013

Selection rules at matched size (E10). Arm A is the MCR subset and arm B is the subset selected by the comparison rule with the same number of features. For identity, a positive change means that arm B has higher identity accuracy. Holm correction is applied across the 86 comparisons in this table.

Table 10: Accuracy of the MCR subset (A) and of each selection rule's subset with the same number of features (B), mean ± SD across subjects (E10). For identity, a positive Δ means that the rule leaks more identity. Δ is B − A in pp, and p is Holm-adjusted across the 86 comparisons in this table.

Dataset	Comparison	n	A (%)	B (%)	Δ (pp)	r	p_{Holm}
ShearSense-Gesture	identity drop target	18	51.3 ± 13.8	50.3 ± 14.3	-1.0	-0.44	1.000
	mutual info target	18	51.3 ± 13.8	52.5 ± 13.1	$+1.2$	0.37	1.000
	pca target	18	51.3 ± 13.8	46.8 ± 14.2	-4.4	-0.71	0.283
	perm target target	18	51.3 ± 13.8	54.4 ± 14.2	$+3.2$	0.75	0.148
	rfe target	18	51.3 ± 13.8	53.1 ± 12.2	$+1.8$	0.47	1.000
	identity drop identity	18	41.7 ± 14.2	40.5 ± 13.3	-1.2	-0.25	1.000
	mutual info identity	18	41.7 ± 14.2	55.0 ± 12.5	$+13.3$	0.98	0.001
	pca identity	18	41.7 ± 14.2	54.0 ± 14.7	$+12.4$	1.00	< 0.001
	perm target identity	18	41.7 ± 14.2	61.3 ± 10.0	$+19.7$	1.00	< 0.001
	rfe identity	18	41.7 ± 14.2	50.6 ± 11.4	$+9.0$	1.00	< 0.001
WISDM (phone accelerometer)	identity drop target	44	28.8 ± 6.5	27.8 ± 7.0	-1.0	-0.29	1.000
	mutual info target	44	28.8 ± 6.5	28.4 ± 7.5	-0.4	-0.08	1.000
	pca target	44	28.8 ± 6.5	24.5 ± 7.5	-4.2	-0.70	0.001
	perm target target	44	28.8 ± 6.5	30.5 ± 7.9	$+1.7$	0.28	1.000
	rfe target	44	28.8 ± 6.5	29.3 ± 7.8	$+0.5$	0.04	1.000
	identity drop identity	44	32.5 ± 12.8	36.0 ± 13.1	$+3.5$	0.29	1.000
	mutual info identity	44	32.5 ± 12.8	87.6 ± 7.6	$+55.1$	1.00	< 0.001
	pca identity	44	32.5 ± 12.8	78.4 ± 10.3	$+45.9$	1.00	< 0.001
	perm target identity	44	32.5 ± 12.8	88.6 ± 6.3	$+56.1$	1.00	< 0.001
	rfe identity	44	32.5 ± 12.8	89.6 ± 6.3	$+57.0$	1.00	< 0.001
WISDM (phone gyroscope)	identity drop target	50	27.4 ± 7.2	25.9 ± 6.7	-1.5	-0.69	0.001
	mutual info target	50	27.4 ± 7.2	31.4 ± 6.6	$+4.0$	0.92	< 0.001
	pca target	50	27.4 ± 7.2	27.1 ± 5.5	-0.3	-0.01	1.000
	perm target target	50	27.4 ± 7.2	31.2 ± 6.6	$+3.8$	0.90	< 0.001
	rfe target	50	27.4 ± 7.2	31.4 ± 6.7	$+4.1$	0.92	< 0.001
	identity drop identity	50	35.7 ± 13.8	37.1 ± 13.6	$+1.4$	0.30	1.000
	mutual info identity	50	35.7 ± 13.8	51.2 ± 14.3	$+15.5$	1.00	< 0.001
	pca identity	50	35.7 ± 13.8	42.4 ± 15.1	$+6.7$	0.72	< 0.001
	perm target identity	50	35.7 ± 13.8	51.3 ± 14.7	$+15.6$	1.00	< 0.001
	rfe identity	50	35.7 ± 13.8	51.6 ± 14.7	$+16.0$	1.00	< 0.001
WISDM (watch accelerometer)	identity drop target	50	52.4 ± 15.3	56.5 ± 16.3	$+4.1$	0.95	< 0.001
	mutual info target	50	52.4 ± 15.3	62.6 ± 16.5	$+10.2$	0.97	< 0.001
	pca target	50	52.4 ± 15.3	58.6 ± 16.4	$+6.1$	0.81	< 0.001
	perm target target	50	52.4 ± 15.3	65.1 ± 15.9	$+12.7$	1.00	< 0.001
	rfe target	50	52.4 ± 15.3	64.2 ± 16.2	$+11.8$	1.00	< 0.001
	identity drop identity	50	24.0 ± 16.5	27.7 ± 16.4	$+3.7$	0.73	< 0.001

Dataset	Comparison	n	A (%)	B (%)	Δ (pp)	r	p_{Holm}
	mutual info identity	50	24.0 ± 16.5	54.6 ± 19.2	$+30.6$	1.00	< 0.001
	pca identity	50	24.0 ± 16.5	50.6 ± 19.4	$+26.6$	1.00	< 0.001
	perm target identity	50	24.0 ± 16.5	59.6 ± 17.4	$+35.6$	1.00	< 0.001
	rfe identity	50	24.0 ± 16.5	55.5 ± 17.7	$+31.4$	1.00	< 0.001
WISDM (watch gyroscope)	identity drop target	50	53.3 ± 15.7	53.3 ± 15.7	0.0	–	–[†]
	mutual info target	50	53.3 ± 15.7	58.5 ± 15.1	$+5.3$	0.98	< 0.001
	pca target	50	53.3 ± 15.7	57.2 ± 15.3	$+3.9$	0.81	< 0.001
	perm target target	50	53.3 ± 15.7	58.5 ± 15.1	$+5.2$	0.98	< 0.001
	rfe target	50	53.3 ± 15.7	58.4 ± 15.0	$+5.1$	0.98	< 0.001
	identity drop identity	50	27.9 ± 11.1	27.9 ± 11.1	0.0	–	–[†]
	mutual info identity	50	27.9 ± 11.1	36.4 ± 12.6	$+8.5$	0.97	< 0.001
	pca identity	50	27.9 ± 11.1	35.3 ± 13.3	$+7.4$	0.84	< 0.001
	perm target identity	50	27.9 ± 11.1	35.8 ± 12.9	$+7.9$	0.94	< 0.001
	rfe identity	50	27.9 ± 11.1	36.3 ± 12.8	$+8.4$	0.97	< 0.001
PAMAP2	identity drop target	9	87.0 ± 9.0	87.0 ± 9.0	0.0	–	–[†]
	mutual info target	9	87.0 ± 9.0	87.0 ± 9.0	0.0	–	–[†]
	pca target	9	87.0 ± 9.0	77.8 ± 20.5	-9.2	-0.51	1.000
	perm target target	9	87.0 ± 9.0	87.0 ± 9.0	0.0	–	–[†]
	rfe target	9	87.0 ± 9.0	87.0 ± 9.0	0.0	–	–[†]
	identity drop identity	9	100.0 ± 0.1	100.0 ± 0.1	0.0	–	–[†]
	mutual info identity	9	100.0 ± 0.1	100.0 ± 0.1	0.0	–	–[†]
	pca identity	9	100.0 ± 0.1	99.1 ± 0.7	-0.9	-1.00	0.328
	perm target identity	9	100.0 ± 0.1	100.0 ± 0.1	0.0	–	–[†]
	rfe identity	9	100.0 ± 0.1	100.0 ± 0.1	0.0	–	–[†]
OPPORTUNITY (locomotion)	identity drop target	4	93.8 ± 4.6	93.8 ± 4.6	0.0	0.33	1.000
	mutual info target	4	93.8 ± 4.6	93.5 ± 4.7	-0.2	-0.60	1.000
	pca target	4	93.8 ± 4.6	91.2 ± 5.3	-2.6	-1.00	1.000
	perm target target	4	93.8 ± 4.6	93.6 ± 5.0	-0.2	0.00	1.000
	rfe target	4	93.8 ± 4.6	93.6 ± 5.1	-0.2	0.20	1.000
	identity drop identity	4	98.9 ± 0.6	99.0 ± 0.4	$+0.1$	0.33	1.000
	mutual info identity	4	98.9 ± 0.6	99.0 ± 0.6	$+0.1$	0.40	1.000
	pca identity	4	98.9 ± 0.6	97.0 ± 1.8	-2.0	-1.00	1.000
	perm target identity	4	98.9 ± 0.6	99.2 ± 0.5	$+0.3$	0.67	1.000
	rfe identity	4	98.9 ± 0.6	99.0 ± 0.6	$+0.1$	0.40	1.000
OPPORTUNITY (both arms)	identity drop target	4	82.9 ± 4.6	82.9 ± 4.6	0.0	–	–[†]
	mutual info target	4	82.9 ± 4.6	82.5 ± 5.3	-0.4	-0.60	1.000
	pca target	4	82.9 ± 4.6	79.7 ± 4.1	-3.2	-1.00	1.000
	perm target target	4	82.9 ± 4.6	82.8 ± 4.9	-0.1	-0.33	1.000
	rfe target	4	82.9 ± 4.6	82.7 ± 4.9	-0.3	-0.67	1.000
	identity drop identity	4	98.6 ± 0.7	98.6 ± 0.7	0.0	–	–[†]
	mutual info identity	4	98.6 ± 0.7	99.0 ± 0.9	$+0.4$	0.67	1.000
	pca identity	4	98.6 ± 0.7	96.1 ± 1.9	-2.5	-1.00	1.000
	perm target identity	4	98.6 ± 0.7	99.1 ± 0.9	$+0.5$	1.00	1.000
	rfe identity	4	98.6 ± 0.7	99.1 ± 0.9	$+0.5$	1.00	1.000
ExtraSensory	identity drop target	30	71.1 ± 9.1	71.1 ± 9.1	0.0	–	–[†]
	mutual info target	30	71.1 ± 9.1	70.8 ± 8.9	-0.3	-0.13	1.000
	pca target	30	71.1 ± 9.1	70.8 ± 8.8	-0.3	-0.31	1.000
	perm target target	30	71.1 ± 9.1	71.0 ± 8.6	-0.1	-0.13	1.000
	rfe target	30	71.1 ± 9.1	70.7 ± 9.1	-0.5	-0.25	1.000
	identity drop identity	30	95.0 ± 5.6	95.0 ± 5.6	0.0	–	–[†]
	mutual info identity	30	95.0 ± 5.6	98.2 ± 2.7	$+3.2$	0.97	0.006
	pca identity	30	95.0 ± 5.6	89.8 ± 7.8	-5.2	-0.92	< 0.001
	perm target identity	30	95.0 ± 5.6	97.9 ± 2.6	$+2.9$	1.00	0.003
	rfe identity	30	95.0 ± 5.6	98.3 ± 2.5	$+3.3$	0.98	0.004
Diabetes 130-US Hospitals	identity drop target	30	53.7 ± 13.8	52.6 ± 12.4	-1.1	-0.25	1.000
	mutual info target	30	53.7 ± 13.8	54.5 ± 12.5	$+0.8$	0.23	1.000
	pca target	30	53.7 ± 13.8	51.8 ± 12.6	-1.9	-0.23	1.000
	perm target target	30	53.7 ± 13.8	57.1 ± 13.2	$+3.4$	0.43	1.000
	rfe target	30	53.7 ± 13.8	54.1 ± 13.1	$+0.4$	0.21	1.000
	identity drop identity	30	72.5 ± 28.5	76.4 ± 24.1	$+3.9$	0.29	1.000
	mutual info identity	30	72.5 ± 28.5	93.9 ± 12.6	$+21.4$	0.91	0.029
	pca identity	30	72.5 ± 28.5	77.8 ± 24.5	$+5.3$	0.28	1.000
	perm target identity	30	72.5 ± 28.5	91.9 ± 13.8	$+19.4$	0.91	0.077
	rfe identity	30	72.5 ± 28.5	93.9 ± 12.6	$+21.4$	0.91	0.029

Differentially private forest (E11). Arm A is a non-private random forest of the same depth and arm B is the differentially private forest at $\varepsilon = 1$, with each subject's accuracy averaged over seeds. Holm correction is applied across the 20 comparisons in this table.

Table 11: Target task and identity accuracy of a non-private forest (A) and a differentially private forest at $\varepsilon = 1$ (B), mean $\pm$ SD across subjects (E11). Δ is B $-$ A in pp, and p is Holm-adjusted across the 20 comparisons in this table.

Dataset	Comparison	n	A (%)	B (%)	Δ (pp)	r	p_{Holm}
ShearSense-Gesture	DP target	18	44.4 ± 14.9	26.6 ± 6.1	-17.8	-1.00	< 0.001
	DP identity	18	27.6 ± 20.8	2.2 ± 4.1	-25.4	-1.00	< 0.001
WISDM (phone accelerometer)	DP target	44	27.4 ± 6.9	17.1 ± 3.8	-10.3	-1.00	< 0.001
	DP identity	44	29.9 ± 29.3	1.5 ± 5.4	-28.4	-1.00	< 0.001
WISDM (phone gyroscope)	DP target	50	27.5 ± 8.5	12.9 ± 2.9	-14.6	-1.00	< 0.001
	DP identity	50	11.4 ± 17.4	0.3 ± 1.3	-11.1	-1.00	< 0.001
WISDM (watch accelerometer)	DP target	50	59.4 ± 18.8	30.4 ± 8.6	-28.9	-0.99	< 0.001
	DP identity	50	10.5 ± 23.4	0.4 ± 2.2	-10.1	-1.00	< 0.001
WISDM (watch gyroscope)	DP target	50	47.3 ± 14.9	19.3 ± 4.6	-28.0	-1.00	< 0.001
	DP identity	50	6.0 ± 11.6	0.2 ± 0.8	-5.7	-0.86	< 0.001
PAMAP2	DP target	9	92.2 ± 4.6	31.0 ± 12.6	-61.2	-1.00	0.027
	DP identity	9	71.9 ± 37.2	0.8 ± 1.1	-71.0	-1.00	0.047
OPPORTUNITY (locomotion)	DP target	4	91.6 ± 5.8	50.4 ± 6.7	-41.2	-1.00	0.625
	DP identity	4	88.3 ± 8.4	2.3 ± 4.7	-86.0	-1.00	0.625
OPPORTUNITY (both arms)	DP target	4	76.6 ± 4.8	38.1 ± 3.4	-38.6	-1.00	0.625
	DP identity	4	92.0 ± 8.5	5.9 ± 6.5	-86.2	-1.00	0.625
ExtraSensory	DP target	30	70.1 ± 10.6	46.8 ± 14.7	-23.3	-1.00	< 0.001
	DP identity	30	76.4 ± 29.0	2.5 ± 5.3	-73.8	-1.00	< 0.001
Diabetes 130-US Hospitals	DP target	30	52.6 ± 13.0	50.0 ± 5.5	-2.6	-0.27	0.625
	DP identity	30	79.2 ± 33.0	3.1 ± 4.5	-76.0	-1.00	< 0.001

D Per-dataset results

This appendix gives the per-dataset detail behind the takeaways of §5, in the order of its subsections.

Features used across near-optimal identity models (E3). Tree identity MCR^{-} is at most 1.000 for every feature on ShearSense, so for each feature at least one tree in the sampled set does not rely on it. The meaning of "every near-optimal identity model" therefore depends on the sampled model class, with 164 GAMs per fold on ShearSense and 275 trees (§5.8).

Near-optimal target task models that avoid the flagged features (E4). On ShearSense, 14 of 18 folds contain no near-optimal target task tree that avoids every flagged feature, and the other 4 do. On WISDM watch accelerometer, no fold contains one (50 of 50). On the phone gyroscope, 32 of 50 folds contain one. On ExtraSensory, which has 3 flagged features, 30 of 30 folds contain such a model. Depth-limited trees use only a subset of the available features, and requiring a tree to avoid all flagged features further restricts the available split variables.

The accuracy–identifiability curve and the cut (E5, E6). On ShearSense, identity accuracy falls at every step of the sweep, from $65.6 \pm 10.1\%$ with all features to $41.7 \pm 14.2\%$ with all 13 flagged features removed. Target task accuracy falls from $55.7 \pm 14.5\%$ to $51.3 \pm 13.8\%$ over the same steps, so the knee is the last step. The curve has a knee on 6 of the 10 datasets, and on 5 of them the knee is the last step, so the cut removes every flagged feature. The knee falls before the last step only on WISDM (watch accelerometer). On OPPORTUNITY (locomotion), OPPORTUNITY (both arms), ExtraSensory, no step lowers identity accuracy by 5 pp, so the cut removes every flagged feature, and PAMAP2 has no flagged feature to remove.

Accuracy at the cut (E7, E9). With the adversary held fixed, identity accuracy on ShearSense changes from $65.6 \pm 10.1\%$ to $41.7 \pm 14.2\%$ (-24.0 pp, Holm $p < 0.001$). Gesture recognition changes from $55.7 \pm 14.5\%$ to $51.3 \pm 13.8\%$ (-4.4 pp). Where the target task change is not a significant cost, it lies between -2.0 and $+0.5$ pp. On ShearSense, TabPFN changes target task accuracy by -4.3 pp, compared with -4.4 pp for the gradient-boosted model, and identity accuracy changes from $73.4 \pm 9.0\%$ to $48.4 \pm 15.0\%$.

Datasets with little change along the removal curve (E5). After removing every flagged feature, identity accuracy is $98.9 \pm 0.6\%$ on OPPORTUNITY locomotion. On ExtraSensory, identity accuracy changes to $95.0 \pm 5.6\%$. Target task accuracy also changes little, including $+0.5$ pp on ExtraSensory.

Matched-size selection rules (E10). In 12 comparisons, the competing rule has significantly higher target task accuracy and significantly higher identity accuracy. In 14, it has significantly higher identity accuracy without a significant target task gain, and in 5 neither difference is significant. In the remaining 2, principal components on ExtraSensory give significantly lower identity accuracy and dropping features by one model's identity importance on WISDM phone gyroscope gives significantly lower target task accuracy, each with no significant difference on the other task. On WISDM watch accelerometer, permutation importance gives identity accuracy of $59.6 \pm 17.4\%$,

compared with $24.0 \pm 16.5\%$ for the flagged subset, with a target task difference of $+12.7$ pp (Holm $p < 0.001$ on both tasks). On ShearSense, principal components give target task accuracy of $46.8 \pm 14.2\%$ and identity accuracy of $54.0 \pm 14.7\%$, compared with $51.3 \pm 13.8\%$ and $41.7 \pm 14.2\%$ for the flagged subset. The selected feature sets have Jaccard similarity from 0.73 to 1.00, with median 0.90.

Sensitivity to the Rashomon-set approximation (E12). Without the varied candidates, trees flag features on PAMAP2 that no tree flags once they are added, and exact enumeration removes the two tree flags of the grid-only sets on the clinical records.

E The cut rule

The rule cuts only flagged features, those whose identity MCR^{-} exceeds 1.02 in every fold for some model class (§4), so the curve it reads removes those features alone. When the curve has no knee, the rule removes every flagged feature, and a dataset without flagged features keeps all of its features. Table 4 gives the number of flagged features at thresholds of 1.01, 1.02 and 1.03. The cut rule can select a different cut from the fixed threshold. The selected point also depends on the rate used to weight identity reduction relative to target task loss.

Let I_k and T_k denote identity and target task accuracy after sweep step k, let I_0 and T_0 denote their values with all features, and let c_I and c_T denote chance accuracy. The rule scores step k by

$$r \, \frac{I_0 - I_k}{I_0 - c_I} \; - \; \frac{T_0 - T_k}{T_0 - c_T}, \tag{2}$$

where the first term is the fraction of above-chance identity accuracy removed and the second is the fraction of above-chance target task accuracy lost. The rate r weights the identity term. We use $r = 1$, giving the two normalized quantities equal weight, and require a reduction in identity accuracy of at least 5 pp before selecting a cut. This score is equivalent to Youden's J (Youden, 1950) on the curve of identity removed against target task lost and selects the last vertex of the lower convex hull before its slope exceeds r. The normalization uses the full feature set and chance as anchors and therefore does not depend on the final point of the sweep.

Table 12 gives, for each dataset, the flagged features, whether the curve has a knee and where it falls, the cut at rates $r = 0.5$, 1 and 2, and how many subjects' cuts, chosen on the other subjects' curve, match the full-curve cut.

Table 12: The cut on each dataset under the per-fold flag. A knee at the last step removes every flagged feature, and a curve without a knee removes every flagged feature as well. The rate columns give the cut when the identity term is weighted by r, and Nested counts the subjects whose cut, chosen on the other subjects' curve, equals the full-curve cut.

				Cut at rate r			
Dataset	Flagged	Knee	Cut	0.5	1	2	Nested
ShearSense-Gesture	13	last step	13	13	13	13	18/18
WISDM (phone accelerometer)	26	last step	26	26	26	26	43/44
WISDM (phone gyroscope)	13	last step	13	3	13	13	50/50
WISDM (watch accelerometer)	28	before the end	27	23	27	28	48/50
WISDM (watch gyroscope)	3	last step	3	2	3	3	50/50
PAMAP2	0	none	0	–	–	–	–
OPPORTUNITY (locomotion)	4	none	4	–	–	–	–
OPPORTUNITY (both arms)	1	none	1	–	–	–	–
ExtraSensory	3	none	3	–	–	–	–
Diabetes 130-US Hospitals	3	last step	3	3	3	3	28/30

F Personalization

We measure how including data from the test person changes target task accuracy on the full and retained feature sets.

For each subject, we hold out 20% of that subject's samples and train on all other subjects plus a fraction f of the subject's remaining samples, where $f \in \{0, 0.1, \ldots, 1\}$. We repeat the experiment over five seeds. With all features, including the person's own data increases target task accuracy significantly on 8 of 10 datasets, by 8.0 to 48.6 pp. With the retained features, it increases target task accuracy significantly on 8 of 10 datasets, by 7.3 to 13.5 pp. App. C reports all datasets.

G Differential privacy

The differentially private forest limits information released about its training set. This guarantee concerns a different quantity from identity information present in the model inputs.

We train diffprivlib's differentially private random forest (Fletcher & Islam, 2017) at several privacy budgets ε and compare it with a non-private forest of the same depth. The feature bounds required by the mechanism are computed from each training split, so the stated privacy budget applies to the forest but not to the preprocessing used to obtain those bounds. Each privacy budget is fitted separately, so the sweep is a sensitivity analysis rather than a single composed release. The library assigns each tree a contiguous subset of rows unless the rows are shuffled. Because our tables are sorted by subject, each tree in the initial run therefore received data from only one or two people. We corrected this by shuffling the rows and reran the phase. With shuffled rows, the forest at $\varepsilon = 1$ reaches target task accuracy of $26.6 \pm 6.1\%$ on ShearSense, compared with $44.4 \pm 14.9\%$ for the non-private forest (Holm $p < 0.001$, Table 11). On the adversarial split, the held-out subject keeps only 80% of their samples in training, so a model without identity signal predicts that subject less often than $1/K$ and scores below chance.

H Window overlap

We test whether overlap between adjacent windows affects identity accuracy on the high-dimensional sensor datasets.

With a one-second step, consecutive windows share samples, so a random adversarial split can place overlapping windows in both training and test sets. We therefore repeat identity evaluation with a blocked split that shares no samples between training and test. Identity accuracy is $97.1 \pm 2.5\%$ on PAMAP2, $97.9 \pm 1.3\%$ on OPPORTUNITY locomotion, and $98.7 \pm 1.2\%$ on OPPORTUNITY both arms (pending the synced tables).

I Reasons to accept an accuracy cost

On 4 of the 10 datasets, the procedure significantly reduces target task accuracy (§5.5). This appendix describes factors that can affect whether a designer accepts that cost. The analysis itself does not require feature removal, and some cuts have small target task costs. Personalization can recover part of the lost target task accuracy on 6 of the 7 datasets where identity accuracy decreases significantly. Other privacy interventions also affect model utility or system design. European data protection law and the AI Act provide additional context for evaluating identity information in model inputs. The final paragraph states the limits of what the procedure changes under these provisions.

Auditing and feature removal are separate decisions. Computing the identity-reliance curve does not require the designer to remove any features. The curve includes the full feature set, where nothing is cut, and reports how well those features support identity prediction. This information can be included in consent materials, documentation, or a risk assessment. The curve also identifies cuts with different target task costs. On WISDM phone accelerometer, identity accuracy changes by -58.0 pp at the cut while target task accuracy changes by -2.0 pp, a difference that the paired test does not distinguish from zero (Holm $p = 0.426$). On the clinical records, identity accuracy changes by -21.4 pp with a target task change of -0.4 pp, although target task accuracy on that dataset is near chance (App. C). On the wide tables of windowed statistics, feature removal produces little change in identity accuracy (§5.6), so the curve also identifies datasets for which this intervention has limited effect.

Personalization can recover target task accuracy. The primary evaluation measures the cost for population models. When training also includes data from the test person, the reduced feature set gives higher target task accuracy than a population model using all features on 6 of the 7 datasets where identity accuracy decreases significantly. The difference is significant on 6 of them in a paired Wilcoxon signed-rank test over subjects, Holm-corrected across all 10 datasets, and App. C reports every dataset. WISDM watch accelerometer is the exception, with $57.9 \pm 13.6\%$ versus $66.3 \pm 16.7\%$, a significant decrease (Holm $p < 0.001$). One deployment option is a shared model using the reduced feature set together with a personal component trained on data that remain on the user's device, as in personalized federated learning (Arivazhagan et al., 2019). Phase E adds the person's samples directly to the training data, so it does not test a separate on-device adapter trained from a short enrollment period. Using all features together with the person's own data gives still higher accuracy in some cases, including $69.3 \pm 11.2\%$ on ShearSense, so personalization does not eliminate the target task cost of feature removal.

Other interventions have different costs and guarantees. Differential privacy bounds what a trained model reveals about an individual training record and can reduce predictive accuracy depending on the privacy budget (Jayaraman & Evans, 2019). Its guarantee does not remove identity information already present in the model inputs (§2). Adversarial encoders instead transform the input representation to reduce information about a sensitive attribute (Malekzadeh et al., 2019). For a dataset release, another option is to withhold features entirely, which also removes the target task information carried by those features from downstream analyses. At $\varepsilon = 1$, the differentially private forest leaves no identity signal and lowers target task accuracy significantly on 7 of 10 datasets (§5.7). These approaches therefore address different quantities and impose different changes on the learning pipeline.

The accuracy and identity effects apply to different parties. The system that uses the reduced feature set incurs the change in target task accuracy. The identity signal concerns the people represented in the data. Identity prediction might be separate from the purpose for which those data were collected. The target task accuracy–identifiability curve reports both quantities to the designer before a cut is selected.

Relevant provisions of the GDPR. Under the GDPR (European Parliament and Council of the European Union, 2016), biometric data are personal data resulting from specific technical processing of a person's physical, physiological, or behavioral characteristics that allow or confirm unique identification (Art. 4(14)). Article 9(1) applies additional restrictions when biometric data are processed for the purpose of uniquely identifying a person. Identity accuracy provides an empirical measure of how well one specified adversary can identify people from the target task features. The GDPR also requires personal data to be adequate, relevant, and limited to what is necessary for their purposes (Art. 5(1)(c)). The target task accuracy–identifiability curve reports how feature removal changes both target task performance and identity prediction. Vertical data minimization addresses a related design problem by controlling how precisely each feature is retained (Staab et al., 2024).

Relevant provisions of the AI Act. Under the AI Act (European Parliament and Council of the European Union, 2024), an emotion recognition system identifies or infers emotions or intentions from biometric data (Art. 3(39)), and Annex III lists specified emotion recognition systems as high-risk (point 1(c)). Providers of high-risk systems must identify and evaluate risks arising under the intended purpose and reasonably foreseeable misuse (Art. 9(2)(b)), limited to risks that can reasonably be mitigated or eliminated through development or design, or through adequate technical information (Art. 9(3)). Identity prediction from the inputs of an affect-recognition system is one potential secondary use that can be evaluated during this process. Feature removal before model training is one possible design intervention for reducing measured identity predictability. Article 9 applies to high-risk systems, so its relevance depends on the system's classification and intended use. The GDPR applies under a different set of conditions to the processing of personal data.

Scope of the procedure. Reducing identity accuracy does not change whether a touch-based affect system satisfies the AI Act's definition of an emotion recognition system, because the Act's definition of biometric data (Art. 3(34)), unlike the GDPR's (Art. 4(14)), does not require that the data allow or confirm a person's unique identification. The Act's definition of a remote biometric identification system concerns systems whose purpose is to identify people without their active involvement, typically at a distance, by comparing their biometric data with the biometric data in a reference database (Art. 3(41)). A gesture or activity model built for another prediction task therefore presents a different use case. Article 5(1)(f) separately prohibits specified emotion-inference uses in workplaces and educational institutions, except for medical or safety reasons. The reduced feature sets in our experiments continue to support identity prediction at 7 to 22 times chance (§5.5). The procedure therefore measures and can reduce empirical identity predictability, but it does not establish that the data are anonymous or that a system falls outside a legal definition. The application of these provisions depends on the specific system and context; our analysis addresses their technical rather than legal interpretation.